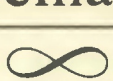
Little Mathematics Library



N.Ya.VILENKIN

METHOD OF SUCCESSIVE APPROXIMATIONS

Mir Publishers. Moscow

ИЗДАТЕЛЬСТВО
МИР

ПОПУЛЯРНЫЕ ЛЕКЦИИ ПО МАТЕМАТИКЕ

Н. Я. Виленкин

МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

LITTLE MATHEMATICS LIBRARY

N.Ya.Vilenkin

METHOD OF SUCCESSIVE APPROXIMATIONS

Translated from the Russian
by
Mark Samokhvalov, Cand. Sc. (Tech.)

MIR PUBLISHERS
MOSCOW

First published 1979

На английском языке

© English translation, Mir Publishers, 1979

Contents

	Preface to the Second Russian Fdition	7
	Preface to the First Russian Edition	8
1	Introduction	9
2.	Successive Approximations	12
3.	Achilles and the Tortoise	14
4	Division On Electronic Computers	16
5	Extraction of Square Roots by Method of Successive Approximations	19
6.	Extraction of Roots with Positive Integer Indices Using Method of Successive Approximations	25
7	Method of Iteration	27
8.	Geometrical Meaning of Method of Iterations	30
9	Contraction Mappings (Contractions)	32
10	Contraction Mappings and Method of Iteration	36
11.	Method of Chords	43
12.	Improved Method of Chords	47
13	Derivative of Polynomial	49
14	Newton's Method for Approximate Solution of Algebraic Equations	51
15.	Geometrical Meaning of Derivative	54
16.	Geometrical Meaning of Newton's Method	57
17	Derivatives of Arbitrary Functions	59
18	Computation of Derivatives	61
19.	Finding the First Approximations	63
20	Combined Method of Solving Equations	66
21	Convergence Test for Method of Iterations	68

22. Rate of Convergence of Iteration Process 71

23. Solving Systems of Linear Equations by Method of Successive Approximations 74

24. Solving Systems of Non-linear Equations Using Method of Successive Approximations 79

25. Modified Distance 82

26. Convergence Tests for Process of Successive Approximations for Systems of Linear Equations 85

27. Successive Approximations in Geometry 91

28 Conclusion 94

Exercises 96

Solutions 98

PREFACE TO THE SECOND RUSSIAN EDITION

For the second edition the book has been revised. The presentation of the method of iteration is now based on the concept of the contraction mapping, as it is possible to consider the latter before introducing the concept of the derivative. The part of the book dealing with the approximate solution of systems of equations has been substantially enlarged. Lastly, all problems have been provided with solutions.

PREFACE TO THE FIRST RUSSIAN EDITION

The main purpose of this book is to present various methods of approximate solution of equations. Their practical value is beyond doubt, but still little attention is paid to them either at school or a college and so someone who has passed a college level higher mathematics course usually has difficulty in solving a transcendental equation of the simplest type. Not only engineers need to solve equations, but also technicians, production technologists and people in other professions as well. It is also good for high-school students to become acquainted with the methods of approximate solution of equations.

Since most approximate solution methods involve the idea of the derivative we were forced to introduce this concept. We did this intuitively, making use of a geometric interpretation. Hence, a knowledge of secondary school mathematics will be sufficient for anyone wanting to read this book.

In writing this book the author made use of a lecture he delivered to 9th and 10th form pupils, members of the school mathematics circle at the Lomonosov State University of Moscow

The material contained in this lecture was used by a teacher at the Moscow secondary school No. 425, S. I. Schwartzburd, for extracurricular work with the 9th form pupils. The author expresses his gratitude to S. I. Schwartzburd for supplying problems involving the solution of equations by the method of iteration. These problems were made use of in the writing of the book.

The author expresses his profound gratitude to V. G. Boltyansky whose remarks were very helpful in improving the original manuscript.

1. Introduction

In studying mathematics at school much time is spent on solving equations and systems of equations Initially equations of the first degree and systems of such equations are studied. Then come quadratic, biquadratic and irrational equations. Finally, the pupil becomes acquainted with exponential, logarithmic and trigonometric equations

It is not by chance that so much attention is paid to equations. The reason is the importance of equations in the practical applications of mathematics. In whatever field of application you choose you will have to solve equations, or systems of equations to arrive at a final answer

At school, equations are often used in solving physics problems Consider, for instance, the following problem.

A stone is thrown into a well. Find the depth of the well if the sound of the stone striking the water is heard T seconds after it has been dropped.

If we denote the depth of the well by x then to find x we obtain the equation

$$\sqrt{\frac{2x}{g}} + \frac{x}{v} = T$$

where v is the sound velocity in air ($\sqrt{2x/g}$ is the time the stone takes to fall, and x/v is the time the sound of the stone striking the water takes to reach us). This is an *irrational equation.* Putting $\sqrt{x} = y$ we reduce it to a quadratic equation

$$\frac{y^2}{v} + \sqrt{\frac{2}{g}}\, y - T = 0$$

which may be solved using the well-known formula

Equations are used to solve geometric problems, as well. For instance, the problem of dividing an interval AB of length l into intervals AC and CB such that $AB : AC = AC : CB$ leads to a quadratic equation

$$x^2 + lx - l^2 = 0$$

where x denotes the length of the interval AC.

The problem of dividing the angle α into three parts leads to a more complex equation. This equation is of the form

$$4x^3 - 3x - \cos\alpha = 0$$

where $x = \cos\alpha/3$ Such equations, called *cubic equations*, are not stu-

died at school, but any course in higher algebra contains the proof that there is a formula for the solution of such equations [see formula (3) below].

However, in physics we often come up against problems which lead to more complex equations, whose solutions are not given either at school, or at university. Take, for instance, an iron bar (the engineers would call it a beam) and fix its ends rigidly. If we strike at the bar, transverse oscillations are generated in it. Mathematical physics shows that to find the frequency of such oscillations the equation

$$e^x + e^{-x} = \frac{2}{\cos x} \tag{1}$$

should be solved, where $e = 2.71828\ldots$.

At school no rules are given for the solution of such equations. Do not think that this is due to the brevity of the school mathematics curriculum. There is *no* formula *at all* for the solution of equation (1) in the sense usually accepted at school. Let's make a more precise statement.

An equation is said to have a solution formula if its roots can be expressed in terms of the parameters of the equation with the aid of the arithmetical operations, the extraction of roots and the exponential, logarithmic, trigonometric and inverse trigonometric functions. In this sense the quadratic equation $x^2 + px + q = 0$ has a solution formula of the form

$$x = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \tag{2}$$

There is a formula for the solution of the cubic equation*

$$x^3 + px + q = 0$$

as well. It is of the form

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \tag{3}$$

However, the use of formula (3) in practice involves a number of difficulties and requires the use of complex numbers.

There is also a formula for the solution of equations of the fourth degree, but it is so complicated that we shall not give it here.

*With the aid of the substitution $x + a_1/3a_0 = y$ any cubic equation $a_0x^3 + a_1x^2 + a_2x + a_3 = 0$ may be reduced to the above form

The situation with the equations of the fifth and higher degree is even worse. The Norwegian mathematician Niels Abel proved in 1826 that for $n \geqslant 5$ there is no formula for the solution of the algebraic equation

$$a_0 x^n + a_1 x^{n-1} + \ \ldots \ + a_n = 0$$

with the aid of arithmetical operations and extraction of roots. Only for particular cases for algebraic equations of a degree higher than the fourth degree are there solution formulas*.

If mathematicians limited their studies to equations having exact solutions, i.e. solutions expressed by some formula, a conversation between an engineer and a mathematician would take the following form.

Engineer When designing a structure I arrived at this equation (shows the equation) I must have a solution quickly — in a month's time I must finish the project

Mathematician I would gladly help you, but there is no solution for an equation of this type

Engineer Couldn't you derive the formula?

Mathematician It's no good trying It has been proved long ago that there is no formula for the solution of such equations

One could imagine that after such a conversation the engineer's opinion of mathematics and its possibilities would change for the worse Happily, such conversations do not take place. Actually, the engineer usually has no need of a formula for the solution of this or that equation. What he needs is an answer with a certain degree of accuracy — whether the answer was obtained from a formula, or by some other means, is not of much interest to him.

Imagine, for instance, that the formula has been found and that the answer obtained from it is $x = 3 + \sqrt{13}$. Clearly, this answer cannot be directly used in practice [one can hardly ask a mechanic to make a part $(3 + \sqrt{13})$ cm long]. For practical purposes $\sqrt{13}$ should be expressed in decimals and as many digits after decimal point should be taken as are required for the given practical problem.

Hence, the engineer will be quite satisfied if the mathematician tells him how to calculate the roots of the equation with the necessary degree of accuracy. Mathematics has developed a number of methods for the approximate solution of equations. Some of them are described in this book.

---

*On algebraic equations see A G Kurosh, *Algebraic Equations of Arbitrary Degrees*, Mir Publishers, Moscow, 1977

2. Successive Approximations

Most methods of approximate solutions of equations are based on the idea of *successive approximations.* This idea is used not only to solve equations, but to solve a number of practical problems, as well.

The method of successive approximations, or the trial and error method, is used by gunners. In order to hit a target they set the azimuth scale and the sight and fire the gun. If the target is missed, the setting of the azimuth scale and of the sight is corrected in accordance with the observed position of the shell's explosion, and the next round is fired. After several approximations they are able to set the azimuth scale and the sight so as to hit the target

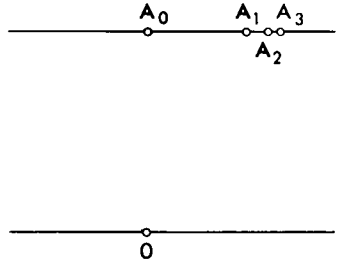


Fig. 1

Sometimes successive approximations are needed also to determine the aiming point. Suppose, an anti-aircraft gun at point O fires at an aircraft in flight (Fig. 1). If the gun is aimed at point A_0 where the plane is at that moment, it will miss, for the plane will move to another point A_1 while the shell travels. If one knows the velocities of the plane and of the shell one may find this point A_1 comparatively easily. However, if the shell is aimed at point A_1 the target may still be missed This is because an inclination of the gun barrel changes the path of the shell's motion and therefore the time taken by the shell to cover the distance OA_0 is not the same as that needed to cover the distance OA_1 and as a result the shell will not hit the plane. But in the latter case the miss will not be as great as when the piece is aimed at point A_0. To make it still less one should find the time taken by the shell to cover the distance OA_1, as well as the point reached by the plane in this time. This point A_2 will be the next approximation for the aiming point sought. After that we shall have to find the time the shell takes to reach point A_2 and calculate the point A_3, which the plane will reach after that time

After several approximations we shall find the aiming point with the necessary degree of accuracy.

The method of successive approximations is also used to solve many other problems.

Suppose one has to transport sand from several sand-pits $A_1, \ldots, A_n$ to several building sites $B_1, \ldots, B_m$. Suppose the productivity of the pit A_j is a_j tons per day, and the amount of sand required by the site B_k is b_k tons per day. Finally, let the cost of transporting a ton of sand from the pit A_j to the site B_k be c_{jk} (this quantity depends on the distance between A_j and B_k, on the state of the roads, etc).

To prepare a transportation plan let us compile Table 1 In this table x_{jk} denotes the amount of sand transported from the pit A_j to the site B_k

Table 1

	B_1	B_2		B_m
A_1	x_{11}	x_{12}		x_{1m}
A_2	x_{21}	x_{22}		x_{2m}
			. .	
A_n	x_{n1}	x_{n2}		x_{nm}

The numbers x_{jk} should of course satisfy the following relations:

$$x_{j1} + x_{j2} + \ . \ . \ + x_{jm} \leqslant a_j$$

(the amount of sand transported from the pit A_j per day should not exceed a_j tons),

$$x_{1k} + x_{2k} + \ \ldots \ + x_{nk} = b_k$$

(the site B_k should receive b_k tons of sand per day).

If the plan given in Table 1 is adopted, the cost of the transportation of sand will be

$$\begin{aligned} c = {} & c_{11}x_{11} + c_{12}x_{12} + \ \ldots \ + c_{1n}x_{1n} + \\ & + c_{21}x_{21} + c_{22}x_{22} + \ \ldots \ + c_{2n}x_{2n} + \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & + c_{m1}x_{m1} + c_{m2}x_{m2} + \ \ldots \ + c_{mn}x_{mn} \end{aligned} \tag{4}$$

The plan should be such that the cost c is the smallest possible. To begin with, a tentative plan is devised. For instance, the following method may be used.

The pit A_1 is made the contractor for the site nearest to it. If the sand production of this pit exceeds the requirements of that site it is made to supply another site, the nearest to it of all the remaining sites. After several steps the productivity of the pit A_1 will be exhausted. After that the pit A_2 is made the contractor for the nearest of the remaining sites, etc. Eventually every site shall have its contractor pit.

However, a plan devised in this way is not the best one because in the end only a few building sites will remain and they may be quite distant from the remaining pits. So the plan will have to be revised. Some of the contracts between the sites and the pits with smaller numbers will have to be cancelled and the sites supplied by pits with greater numbers.

The methods of changing the plan, leading to a reduction in transportation costs, are considered in a branch of mathematics called *linear programming* *.

After several approximations made with the aid of these methods we shall arrive at a plan for which the sum (4) is a minimum, or close to a minimum.

In general, when devising a plan, a timetable, etc. the practice is to begin with some rough approximation which is subsequently improved successively until finally the required result is obtained.

The machining of some part in a factory work shop may also be considered to be a process of successive approximation to the desired shape. At first some rough approximation—a casting or a blank is taken. This blank is machined on a lathe to obtain a shape close to that of the part being fabricated. After that it is passed over to a more accurate lathe. The required part is produced after several such machining operations, i.e. after several approximations.

3. Achilles and the Tortoise

The first to mention successive approximations was Zeno of Elea who lived *c.* 500 B. C. This philosopher tried to prove that there was no motion in nature. The following reasoning was used by Zeno to prove the absence of motion: if the fastest Greek runner Achilles were

---

* On linear programming see A S Solodovnikov "*Introduction to linear algebra and linear programming*", Prosveshcheniye, Moscow, 1966

to try to catch up with a tortoise, he would be unable to. Indeed, suppose the distance between Achilles and the tortoise is 1000 steps, and in one second Achilles runs 10 steps, while the tortoise crawls one step. In 100 seconds Achilles will run 1000 steps separating him from the tortoise. But during this time the tortoise will crawl 100 steps. In 10 seconds Achilles will run 100 steps, but the tortoise will crawl a further 10 steps. To cover this distance Achilles will need another second during which the tortoise will move one step further. Hence, the tortoise will always be in front of Achilles and he will never be able to catch up with it. Consequently, there is no motion.

Of course, such reasoning by Zeno is only a witty paradox and nothing more. Motion is an intrinsic property of matter.

Any pupil will have no difficulty in calculating when Achilles catches up with the tortoise. To do this one should formulate an equation

$$10x - x = 1000 \tag{5}$$

where x is the time sought. From this equation we obtain

$$x = \frac{1000}{9} \text{ s} = 111\frac{1}{9} \text{ s}$$

where 's' denotes seconds.

However, Zeno's reasoning may be regarded as a particular method of approximate solution of equation (5).

Indeed, transpose x into the right-hand side of the equation and divide both sides by 10. We shall obtain the equation

$$x = 100 + \frac{x}{10} \tag{6}$$

If in the right-hand side we neglect the term $x/10$ (it is small compared with x) we obtain for x the approximate solution $x_1 = 100$. Now we can make the answer more precise by substituting for x in the right-hand side the approximation obtained, $x_1 = 100$. We shall obtain a more accurate value for x, i.e. $x_2 = 100 + 10 = 110$. Substituting the new value into the right-hand side of the equation we find the next approximation $x_3 = 100 + 110/10 = 111$. In this way we obtain the approximations

$$x_1 = 100,\ x_2 = 110,\ x_3 = 111,\ x_4 = 111.1,\ \ldots$$

i.e. the same numbers that followed from Zeno's reasoning. These

numbers are connected by the following relationship

$$x_{n+1} = 100 + \frac{x_n}{10} \tag{7}$$

which enables successive calculations of them to be made. As n increases they approach the exact solution $x = 111\frac{1}{9}$ of equation (5).

The method of solution described above proved successful because the term $x/10$ was small in comparison with x. Otherwise we would have obtained numbers which did not get closer and closer to the solution sought Suppose, for instance, that Achilles took on not a slow tortoise, but a light-footed antelope, which runs 20 steps per second. To find the time it will take Achilles to catch up with the antelope we must solve the equation

$$10x - 20x = 1000 \tag{8}$$

Its solution is $x = -100$. This means that Achilles and the antelope ran neck and neck 100 seconds ago, and that now the antelope has overtaken Achilles, the distance separating them growing with time.

Let us try to solve equation (8) using the same method we used to solve equation (5). To do this transpose the term $20x$ into the right-hand side and divide both sides of the equation by 10. We obtain the equation

$$x = 100 + 2x \tag{9}$$

Put $x_0 = 0$ in the right-hand side We find that $x_1 = 100$. Substituting this value into the right-hand side of equation (9) we obtain the next approximation $x_2 = 300$. Continuing the process we obtain the numbers

$$x_0 = 0,\ x_1 = 100,\ x_2 = 300,\ x_3 = 700,\ \ldots$$

We see that the numbers do not approach the exact solution $x = -100$ of equation (8).

4. Division on Electronic Computers

The reader may perhaps be puzzled: why solve equation (5) by the method of successive approximations when it is simple to solve it exactly. But of course equation (5) itself was of little interest to us — we were interested in the method of successive approximations which we intend to apply to more complex equations.

By the way, it is worth mentioning that with the appearance of high speed electronic computers it has become necessary fairly recently to solve equations similar to Eq. (5) by means of the method of successive approximations. Some computers can perform only three arithmetical operations: addition, subtraction and multiplication. Apart from that they can divide by numbers of the form 2^n. What is the method such computers use to divide by arbitrary numbers?

The division of the number b by the number a entails the solution of the equation $ax = b$. Since the computer can multiply and divide by 2^n we may assume that $1/2 \leqslant a < 1$ (otherwise we can multiply, or divide, both sides of the equation $ax = b$ by the number 2 raised to the appropriate power). Rewrite the equation $ax = b$ in the form

$$x = (1 - a)x + b \tag{10}$$

Let $x_1 = b$ be the first approximation for x. Denote the error of this approximation by α_1, i.e. suppose that $x_1 + \alpha_1 = b/a$. Then we obtain from equation (10)

$$\begin{aligned} x_1 + \alpha_1 &= (1 - a)(x_1 + \alpha_1) + b = \\ &= (1 - a)x_1 + b + (1 - a)\alpha_1 \end{aligned} \tag{11}$$

Since $1/2 \leqslant a < 1$ it follows that

$$0 < 1 - a \leqslant 1/2$$

The factor $1 - a$ being comparatively small we discard the term $(1 - a)\alpha_1$ in the right-hand side of equation (11) which is not greater than $\alpha_1/2$. We obtain

$$x_1 + \alpha_1 \approx (1 - a)x_1 + b$$

The number

$$x_2 = (1 - a)x_1 + b$$

will be taken as the next approximation for x.

Denote the error of the approximation x_2 by α_2, i.e. put $x_2 + \alpha_2 = b/a$. Then we obtain from equation (10)

$$x_2 + \alpha_2 = (1 - a)x_2 + b + (1 - a)\alpha_2$$

Discarding the term $(1 - a)\alpha_2$ in the right-hand side of this equation we obtain the approximate equation

$$x_2 + \alpha_2 \approx (1 - a)x_2 + b$$

Hence, we may choose as the next approximation

$$x_3 = (1 - a)x_2 + b$$

By similar reasoning we arrive at the next approximation

$$x_4 = (1 - a)x_3 + b$$

etc. The numbers $x_1, x_2, \ldots, x_n, \ldots$ successively computed from the formula

$$x_{n+1} = (1 - a)x_n + b \tag{12}$$

approach the number b/a. But this formula makes use only of the operations of addition, subtraction and multiplication, and this means that the computer may use it for calculating.

The division method described above is actually based on the formula for the sum of an infinite decreasing geometric progression. Indeed, writing the fraction b/a in the form

$$\frac{b}{a} = \frac{b}{1 - (1 - a)}$$

we obtain with the above-mentioned formula

$$\frac{b}{1 - (1 - a)} = b + b(1 - a) + b(1 - a)^2 + \ldots + b(1 - a)^{n-1} + \ldots \tag{13}$$

Denote the sum of the first n terms of this progression by x_n,

$$x_n = b + b(1 - a) + \ldots + b(1 - a)^{n-1}$$

Obviously

$$\begin{aligned} x_{n+1} &= b + b(1 - a) + \ldots + b(1 - a)^n = \\ &= b + (1 - a)[b + b(1 - a) + \ldots \\ &\ldots + b(1 - a)^{n-1}] = b + (1 - a)x_n \end{aligned}$$

This formula coincides with formula (12). Hence, by substituting the approximate value x_n for the fraction b/a we substitute the sum of the first n terms for the infinite sum in formula (13). As the number n of summands increases, this sum approaches the sum of the entire progression (the progression (13) is a decreasing one since $1/2 \leqslant a < 1$ and so $0 < 1 - a \leqslant 1/2$).

5. Extraction of Square Roots by Method of Successive Approximations

Let's demonstrate now how the method of successive approximations is applied to the extraction of square roots. At school a method is learned which enables the decimal digits of a square root to be found one after the other. It, too, may be regarded as being a method of successively approximating the answer However, this method is rather complicated, and students often use it mechanically not fully understanding how it works. We shall describe another method which was in use in ancient Babylon. It was also used by the Greek geometer Hero (Heron) of Alexandria. Subsequently this method was forgotten, but now it is sometimes used in electronic computers to extract square roots.

Suppose, for instance, we have to extract the square root of the number 28. At first choose some approximate value of this root, for instance, put $x_1 = 5$ We shall denote the error of this approximate value by α_1, i.e. we shall put $\sqrt{28} = 5 + \alpha_1$. To obtain α_1 we take the square of both sides of the equation and obtain

$$28 = 25 + 10\alpha_1 + \alpha_1^2$$

i e.

$$\alpha_1^2 + 10\alpha_1 - 3 = 0 \tag{14}$$

Thus, we have obtained a quadratic equation for α_1. If we try to solve this equation exactly, we obtain $\alpha = -5 \pm \sqrt{28}$. Hence, to find α_1 accurately we must compute $\sqrt{28}$. It seems that we have found ourselves in a vicious circle: to find $\sqrt{28}$ we must compute α_1 and to compute α_1 we must calculate $\sqrt{28}$.

The following reasoning comes to our rescue. The error α_1 in the approximate value $x_1 = 5$ is not large, certainly less than unity. The number α_1^2 is still less. Therefore we shall try to find α_1 discarding the small term α_1^2 in equation (14). Then we shall obtain for α_1 the approximate equation $10\alpha_1 - 3 \approx 0$ whence $\alpha_1 \approx 0.3$.

Thus, we have found the approximate value of the correction α_1 Since $\sqrt{28} = 5 + \alpha_1$, the second approximation x_2 for $\sqrt{28}$ takes the form

$$x_2 = 5 + 0.3 = 5.3$$

To obtain a still more accurate approximation for $\sqrt{28}$ let us repeat the above process, i.e. let us denote the error in the value $x_2 = 5.3$ by α_2 putting $\sqrt{28} = x_2 + \alpha_2$ Taking the squares of both sides of this equality and discarding the small term α_2^2 we get $28 \approx x_2^2 + 2x_2\alpha_2$ and therefore

$$\alpha_2 \approx \frac{28 - x_2^2}{2x_2}$$

This means that the formula for the third approximation for $\sqrt{28}$ is

$$x_3 = x_2 + \frac{28 - x_2^2}{2x_2} = \frac{28 + x_2^2}{2x_2}$$

Since $x_2 = 5.3$ we obtain from here $x_3 = 5.2915...$ In the same way starting with the approximate value $x_3 = 5.2915$, we obtain the next approximation x_4 expressed by the formula

$$x_4 = \frac{28 + x_3^2}{2x_3} = 5\,2915$$

Generally, if we have already found the approximation x_n for $\sqrt{28}$ the next approximation for it is

$$x_{n+1} = \frac{28 + x_n^2}{2x_n} \tag{15}$$

Thereby every new step in the process gives us ever more accurate approximations for $\sqrt{28}$ The computation process stops when the difference between x_{n+1} and x_n becomes less than the specified computation accuracy. For instance, if we have to compute $\sqrt{28}$ with an accuracy up to 0.0001, four approximations are enough and we may put $\sqrt{28} = 5\,2915$ (indeed, $x_3 = 5.2915...$ and $x_4 = 5.2915...$)

The same method may be used to extract the square root of any other positive number. Thus, when computing $\sqrt{a}$ we choose some initial approximate value x_1 and then compute the next approximations with the aid of the formula

$$x_{n+1} = \frac{a + x_n^2}{2x_n} \tag{16}$$

Formula (16) may be derived by reasoning somewhat different than that used in extracting the roof of $\sqrt{28}$. Suppose we have already found the n-th approximation x_n for $\sqrt{a}$. Since $\sqrt{a} = \sqrt{x_n \cdot \frac{a}{x_n}}$ it follows that $\sqrt{a}$ is the geometric mean of the numbers x_n and $\frac{a}{x_n}$. We shall take the approximate value of this geometric mean as the

arithmetic mean of the numbers x_n and $\frac{a}{x_n}$, i.e. we shall put

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) = \frac{x_n^2 + a}{2x_n}$$

This is just formula (16).

Hence, *the method of approximate extraction of square roots described above consists in substituting the arithmetic mean for the geometric mean of the numbers* x_n *and* $\frac{a}{x_n}$ *at every step.*

Let us now discuss whether or not the process of successive approximations as applied to the extraction of square roots always leads to an answer, i.e. whether the situation is always the same as in the case of Achilles and the tortoise, or whether it is sometimes as in the case of Achilles running after the antelope (mathematicians say that in the first case the process converges and in the second case diverges). We shall prove that the process of extracting square roots is never fraught with complications—it is always a convergent process and it always leads to the desired result.

To this end let us compare the errors $\alpha_n = \sqrt{a} - x_n$ and $\alpha_{n+1} = \sqrt{a} - x_{n+1}$ of two successive approximations. The error α_{n+1} may in accordance with formula (16) be written in the form

$$\alpha_{n+1} = \sqrt{a} - x_{n+1} = \sqrt{a} - \frac{x_n^2 + a}{2x_n} = -\frac{x_n^2 - 2x_n\sqrt{a} + a}{2x_n}$$

But

$$x_n^2 - 2x_n\sqrt{a} + a = (x_n - \sqrt{a})^2 = \alpha_n^2$$

and therefore

$$\alpha_{n+1} = -\frac{\alpha_n^2}{2x_n} \tag{17}$$

We consider only positive approximate values x_n of $\sqrt{a}$. Therefore we may draw the conclusion from equality (17) that all the errors $\alpha_2, \alpha_3, \ldots, \alpha_n, \ldots$ are negative. In other words, *all approximations starting with the second one are excessive approximations* *; the first approximation x_1 may be either excessive, or deficient.

*The explanation is that the arithmetic mean is always greater than the geometric mean.

With the aid of formula (17) it may easily be proved that *the absolute value of the error in the approximate value x_n decreases at least twice with every step.* Indeed, equality (17) may be written down in the form

$$\alpha_{n+1} = -\frac{\alpha_n}{2x_n}\alpha_n = \frac{x_n - \sqrt{a}}{2x_n}\alpha_n = \left(\frac{1}{2} - \frac{\sqrt{a}}{2x_n}\right)\alpha_n$$

Therefore

$$|\alpha_{n+1}| = \left|\frac{1}{2} - \frac{\sqrt{a}}{2x_n}\right||\alpha_n| \tag{18}$$

But since $x_n > 0$ it follows that

$$\frac{1}{2} - \frac{\sqrt{a}}{2x_n} < \frac{1}{2}$$

On the other hand, as was shown above, for $n \geqslant 2$ we have $x_n > \sqrt{a}$ and therefore

$$\frac{1}{2} - \frac{\sqrt{a}}{2x_n} > 0$$

This leads to the inequality

$$\left|\frac{1}{2} - \frac{\sqrt{a}}{2x_n}\right| < \frac{1}{2} \tag{19}$$

Comparing relations (18) and (19) we see that

$$|\alpha_{n+1}| < \frac{1}{2}|\alpha_n|$$

This proves the truth of our statement: with every step the absolute value of the error decreases to less than half its previous value. This means that after the second approximation step the error will decrease in absolute value to less than one quarter of its original value, after the third — to less than one eighth, etc. Clearly, as n increases, the absolute value of the error $\alpha_n = \sqrt{a} - x_n$ will decrease and tend to zero. But this just means that the numbers x_n tend to $\sqrt{a}$ as n increases.

Let us now discuss how the choice of the initial approximation x_1 affects the approximation process. To begin with note that this choice has no effect whatsoever on the final result for we have already proved that no matter what initial approximation x_1 was chosen, the errors $\alpha_2, \ldots, \alpha_n, \ldots$ of the subsequent approximations tend to zero as $n \to \infty$. Hence, if the necessary computation accuracy is specified, the same

value of $\sqrt{a}$ within that accuracy will be obtained for all initial approximations x_1. Even if the choice of the initial approximation is made very badly, we shall eventually arrive at the correct result. After ten approximation steps the absolute value of the error will decrease at least a thousand times ($2^{10} = 1024 \approx 1000$) and after forty — at least a billion (10^{12}) times. Thus, if when computing $\sqrt{2}$ we put $x_1 = 10^6$ so that $\alpha_1 \approx 10^6$, then $|\alpha_{40}| < 10^{-6}$. In other words, in the beginning of the process the error was about a million, and at the end its absolute value became less than one millionth.

Nevertheless, the choice of the initial approximation affects the length of the approximation process. If the initial approximation is unfortunate, one has to wait a long time before the difference between x_{n+1} and x_n becomes less than the specified computation accuracy. A good choice of the initial approximation speeds up the process. Hence it is often the practice to take the initial approximation from the tables of square roots and to use the formula

$$x_2 = \frac{a + x_1^2}{2x_1} \tag{20}$$

only to obtain a more precise value.

This method is especially convenient because the rate of decrease in the error is appreciably higher as x_n approaches $\sqrt{a}$. This is because in deriving the inequality

$$|\alpha_{n+1}| < \frac{1}{2}|\alpha_n|$$

we have substituted the number 1/2 for the factor $\left|\frac{1}{2} - \frac{\sqrt{a}}{2x_n}\right|$ in formula (18). However, if x_n is close to $\sqrt{a}$, the fraction $\frac{1}{2} - \frac{\sqrt{a}}{2x_n}$ is very small and therefore $|\alpha_{n+1}| = \left|\frac{1}{2} - \frac{\sqrt{a}}{2x_n}\right||\alpha_n|$ is much less than $\frac{1}{2}|\alpha_n|$.

We can say this more precisely. To do so consider together with the absolute error $|\alpha_n| = |\sqrt{a} - x_n|$, the relative error β_n of the approximate value x_n, i.e. the ratio of the absolute error $|\alpha_n|$ to the exact value of the root $\sqrt{a}$. This error is expressed by the formula

$$\beta_n = \frac{|\alpha_n|}{\sqrt{a}} = \left|1 - \frac{x_n}{\sqrt{a}}\right|.$$

The following formula for the quantity β_{n+1} may be obtained from equation (17):

$$\beta_{n+1} = \frac{|\alpha_{n+1}|}{\sqrt{a}} = \frac{|\alpha_n|^2}{2x_n\sqrt{a}}$$

Since $x_n > \sqrt{a}$ it follows that

$$\beta_{n+1} < \frac{|\alpha_n|^2}{2(\sqrt{a})^2} = \frac{1}{2}\beta_n^2$$

Thus, the relative errors β_n satisfy the inequality

$$\beta_{n+1} < \frac{\beta_n^2}{2} \tag{21}$$

For instance, if the relative error of the approximation x_n is 0.01 it does not exceed 0.000 05 for x_{n+1} and 0.000 000 000 13 for x_{n+2}. We see that accuracy of the approximations improves at an ever increasing rate. It may be demonstrated that when we are quite close to $\sqrt{a}$ every successive approximation doubles the number of correct significant digits.

Example. Compute $\sqrt{238}$ with an accuracy of 0.00001.

From a table of square roots we find $\sqrt{238} = 15.43$. Put $x_1 = 15.43$ and find x_2 using the formula

$$x_2 = \frac{15\,43^2 + 238}{30.86} = 15.42725.$$

Assess the accuracy of the answer obtained. Since the error of the value 15.43 does not exceed 0.01, $\alpha_1 = 0.01$ and therefore

$$\beta_1 \approx \frac{0.01}{15.43} < 0.001$$

But in this case

$$\beta_2 < \frac{0.001^2}{2} = 0.000\,0005$$

This means that the absolute error of the approximation x_2 does not exceed the value $15.43 \times 0.000\,0005 < 0.000\,01$. In other words, all seven digits of the value $\sqrt{238} = 15.427\,25$ are correct.

If we wanted to have fourteen correct digits we could obtain the necessary result from just the third approximation. However, the need for such accuracy is very rare.

Let us finally mention the following peculiarity of the method of successive approximations When the usual method of extracting square roots is employed, an error made at any stage completely invalidates all subsequent computations. The situation is different when the method of successive approximations is used. Suppose that as a result of an error we obtained a wrong value y_n of the n-th approximation instead of the right value x_n. In this case all the subsequent computations may be regarded as computations of $\sqrt{a}$ with the initial approximation y_n. But we have already seen above that the method of successive approximations leads us to the correct value of $\sqrt{a}$ to the required accuracy no matter what initial value was chosen. Hence, the error we made will eventually tend to zero. The only effect it will have is to force us to take a few extra approximation steps.

Because of this peculiarity of the method of successive approximations the computations may be started with low accuracy, the specified accuracy being employed only for the final approximations. This shortens the time needed for the computations.

6. Extraction of Roots with Positive Integer Indices Using Method of Successive Approximations

The method of extracting square roots described above may be used for extracting roots with any positive integer index, as well. For this purpose we shall need the formula *

$$(x+\alpha)^k = x^k + kx^{k-1}\alpha + \ \ldots \tag{22}$$

where the dots denote terms containing α^2, α^3, etc.

Let us prove this formula. It is known from the school mathematics course that

$$(x+\alpha)^2 = x^2 + 2x\alpha + \alpha^2$$

$$(x+\alpha)^3 = x^3 + 3x^2\alpha + 3x\alpha^2 + \alpha^3.$$

These equations may be re-written in the following form:

$$(x+\alpha)^2 = x^2 + 2x\alpha + \ \ldots \tag{23}$$

$$(x+\alpha)^3 = x^3 + 3x^2\alpha + \ \ldots \tag{24}$$

*This formula follows from the binomial theorem, but we do not expect the reader to be acquainted with this theorem

Hence, the formula (22) has been proved for $k = 2$ and $k = 3$. Multiply now both sides of formula (24) by $(x + \alpha)$. We shall obtain that

$$(x + \alpha)^4 = (x^3 + 3x^2\alpha + \ \ldots)(x + \alpha)$$

If we remove the brackets in this equation we obtain one term x^4 not containing α, and two terms $3x^3\alpha$ and $x^3\alpha$ containing α to the first power; the other terms containing α to the second and higher powers. Therefore we may write

$$(x + \alpha)^4 = x^4 + 3x^3\alpha + x^3\alpha + \ \ldots \ = x^4 + 4x^3\alpha + \ \ldots \quad (25)$$

(as before, the dots denote terms containing α^2, α^3, etc.).

Thus, formula (22) has been proved for $k = 4$, as well. In the same way formula (25) yields

$$(x + \alpha)^5 = x^5 + 5x^4\alpha + \ \ldots \quad (26)$$

Obviously, in the same way we may prove formula (22) for any positive integer exponent k.

Let us now return to the extraction of a k-th root, where k is any whole number. Suppose that some approximation x_1 for the sought root $\sqrt[k]{a}$ has been found. Denote the error of this approximation by α_1, i.e. suppose that $x_1 + \alpha_1 = \sqrt[k]{a}$. Then $(x_1 + \alpha_1)^k = a$ But using formula (22) we may write this equation in the following form:

$$x_1^k + kx_1^{k-1}\alpha_1 + \ \ldots = a$$

where the dots denote terms containing α_1^2, α_1^3, etc.

If the approximation x_1 chosen was close enough to $\sqrt[k]{a}$, the error α_1 of this approximation will be small and we will be able to neglect terms containing higher powers of the error. Hence, we obtain the following approximate equality:

$$x_1^k + kx_1^{k-1}\alpha_1 \approx a$$

It follows from this equality that

$$\alpha_1 \approx \frac{a - x_1^k}{kx_1^{k-1}}$$

and for this reason we may take as the next approximation for $\sqrt[k]{a}$ the number

$$x_2 = x_1 + \frac{a - x_1^k}{kx_1^{k-1}} = \frac{a + (k-1)x_1^k}{kx_1^{k-1}}$$

In the same way, using the approximation x_2, we may find the next

approximation

$$x_3 = \frac{a + (k-1)x_2^k}{kx_2^{k-1}}$$

In general, if the approximation x_n for $\sqrt[k]{a}$ has been found, the next approximation will be given by the formula

$$x_{n+1} = \frac{a + (k-1)x_n^k}{kx_n^{k-1}} \tag{27}$$

As was the case with the extraction of square roots, it may be shown that the above process converges for any initial approximation x_1 (provided this approximation is a positive number). In other words, for any x_1 chosen, the numbers $x_1, x_2, \ldots, x_n, \ldots$ tend to $\sqrt[k]{a}$. The approximation process is continued until the numbers x_n and x_{n+1} coincide within the accuracy required.

Example. Find the value of $\sqrt[3]{970}$ to an accuracy of 0.001. For $k = 3$ the approximation formula (27) assumes the form

$$x_{n+1} = \frac{a + 2x_n^3}{3x_n^2} \tag{28}$$

In our case $a = 970$. Put $x_1 = 10$. It follows from formula (28) that

$$x_2 = \frac{970 + 2 \times 10^3}{3 \times 10^2} = \frac{2970}{300} = 9.900$$

$$x_3 = \frac{970 + 2 \times 9.9^3}{3 \times 9.9^2} = \frac{2910.60}{294.03} = 9.899$$

We see that the values of x_2 and x_3 coincide within the accuracy specified. Therefore we have with an accuracy of 0.001

$$\sqrt[3]{970} = 9.899$$

7. Method of Iteration

All the examples considered above are specific cases of a single general method of solving equations. This method is called the *method of iteration*, or the *method of successive approximations*. The essence of this method is as follows.

The equation $f(x) = 0$ which is to be solved is rewritten in the form

$$x = \varphi(x) \tag{29}$$

Then an initial approximation x_1 is chosen and substituted into the right-hand side of (29). The value $x_2 = \varphi(x_1)$ so obtained is taken as the second approximation for the root. In general, if the approximation x_n has been found, the next approximation x_{n+1} is obtained from the formula

$$x_{n+1} = \varphi(x_n)$$

Suppose that after several approximations the equality $x_n \approx x_{n+1}$ is satisfied within the specified accuracy. Since $x_{n+1} = \varphi(x_n)$ this means that the equation $x_n \approx \varphi(x_n)$ is also satisfied within that accuracy, i.e that x_n is the approximate value of the root of the equation $x = \varphi(x)$.

For instance, in solving the problem of Achilles and the tortoise we wrote the equation

$$10x - x = 1000$$

in the form

$$x = 100 + \frac{x}{10}$$

and looked for approximations in the form

$$x_{n+1} = 100 + \frac{x_n}{10}$$

In the problem concerning division on an electronic computer we wrote down the equation

$$ax = b$$

in the form

$$x = (1 - a)x + b$$

and looked for approximations given by the formula

$$x_{n+1} = (1 - a)x_n + b$$

Finally, when extracting the k-th root we transformed the equation

$$x^k = a$$

into

$$x = \frac{a + (k-1)x^k}{kx^{k-1}}$$

after which we looked for approximations using the formula

$$x_{n+1} = \frac{a + (k-1)x_n^k}{kx_n^{k-1}}$$

Here is an example of a more complex equation which can be solved by the method of iterations.

Example. Solve the equation

$$10x - 1 - \cos x = 0 \tag{30}$$

with an accuracy of 0.001.

Rewrite equation (30) in the following form:

$$x = \frac{1 + \cos x}{10} \tag{31}$$

Choose some initial approximation, for instance $x_1 = 0$, and substitute it into the right-hand side of equation (31). The value obtained

$$x_2 = \frac{1 + \cos 0}{10} = 0.2$$

will serve as the second approximation for the root sought. Substituting the value of x_2 into the right-hand side of equation (31) we obtain the third approximation:

$$x_3 = \frac{1 + \cos 0.2}{10} \approx \frac{1 + 0.98}{10} = 0.198.$$

Next we find

$$x_4 = \frac{1 + \cos 0.198}{10} \approx 0\,198$$

We see that the equality $x_3 = x_4$ is satisfied with an accuracy of 0.001. Since $x_4 = \dfrac{1 + \cos x_3}{10}$ this means that, to an accuracy of 0.001, the number $x_3 = 0.198$ is the root of the equation $x = \dfrac{1 + \cos x}{10}$.

Several questions arise in connection with the method of iterations:

1. Does the sequence $x_1, \ldots, x_n, \ldots$ obtained by the iterative method always converge to some number ξ?

2. If the equality $\lim\limits_{n \to \infty} x_n = \xi$ is true, is the number ξ a solution of the equation $x = \varphi(x)$?

3. How rapidly do the numbers $x_1, \ldots, x_n, \ldots$ approach the root of the equation $x = \varphi(x)$?

The second question is the easiest to answer. Suppose the numbers $x_1, \ldots, x_n, \ldots$ approach the number ξ. Consider the equality $x_{n+1} =$

$= \varphi(x_n)$ which expresses the next approximation in terms of the preceding one. As n increases, its left-hand side approaches ξ and the right-hand side approaches $\varphi(\xi)$*. Hence we obtain in the limit $\xi = \varphi(\xi)$, i.e. ξ is the root of the equation $x = \varphi(x)$

The answer to the first question is in the negative. Indeed, consider, for instance, the equation

$$x = 10^x - 2$$

If we put $x_1 = 1$ we obtain

$$x_2 = 8, \quad x_3 = 10^8 - 2, \ldots$$

As n increases, the numbers x_n also increase, but do not tend to any limit. On the other hand, if we rewrite the equation in the form $x = \log(x + 2)$, the approximation process will converge and we obtain after three approximations $x = 2.38$.

Therefore instead of the first question we should ask the following one:

What form of the function $\varphi(x)$ guarantees the convergence of the sequence of numbers $x_1, x_2, \ldots, x_n, \ldots$?

Before dealing with this question we shall discuss the geometrical interpretation of the method of iterations.

8. Geometrical Meaning of Method of Iterations

Clearly, finding the root ξ of the equation $x = \varphi(x)$ is just the same as finding the abscissa of the point M of intersection of the curve $y = \varphi(x)$ with the straight line $y = x$. Suppose we have some initial value x_1 (Fig. 2). In this case the point M_1 with the coordinates $M_1(x_1, \varphi(x_1))$ lies on the curve $y = \varphi(x)$. Draw a horizontal line through this point. It will intersect the straight line $y = x$ in the point $N_1(\varphi(x_1), \varphi(x_1))$. Denote $\varphi(x_1)$ by x_2. Then the coordinates of the point N_1 will be of the form $N_1(x_2, x_2)$. Next draw a vertical line through the point N_1. It will intersect the curve $y = \varphi(x)$ in the point M_2 with the coordinates $M_2(x_2, \varphi(x_2))$. Repeating the process, we obtain the point N_2 on the straight line $y = x$ with the coordinates $N_2(x_3, x_3)$ where $x_3 = \varphi(x_2)$, then the point M_3 on the curve $y = \varphi(x)$ with the coordinates $M_3(x_3, \varphi(x_3))$, etc. If the approximation process converges, the points $M_1, M_2, \ldots, M_n, \ldots$ will approach the point of intersection sought.

*We suppose $\varphi(x)$ to be a continuous function

Hence, the geometrical meaning of the method of successive approximations is that we move towards the required point of intersection of the curve and the straight line along a broken line whose vertices lie in turn on the curve and on the straight line and whose sections are in turn horizontal and vertical (Fig. 2*a*).

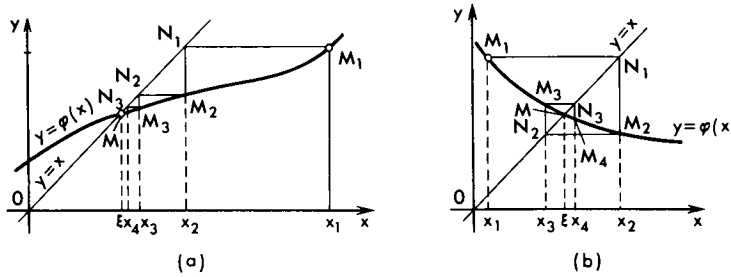


Fig 2

If the curve and the straight line are situated as shown in Fig. 2*a*, then this broken line looks like a ladder. If, on the other hand, the curve and the straight line are as shown in Fig. 2*b*, then the broken line looks like a spiral.

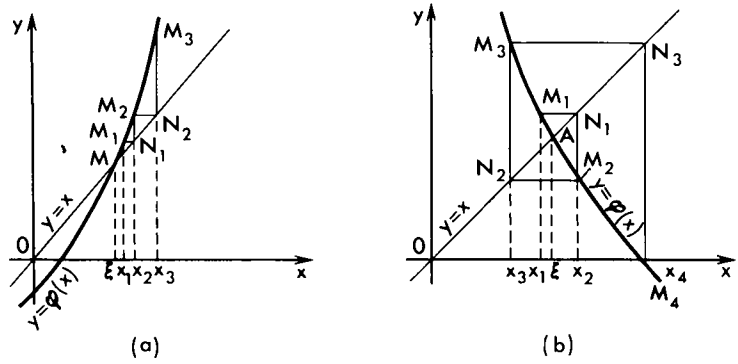


Fig 3

The process of successive approximations described above may diverge without leading to any result (as was so in the case of the problem of Achilles and the antelope). Graphically it means that the steps of the ladder (or the spiral) become larger and larger and because of this the points M_1, ..., M_n, ... instead of approaching the point M move away from it (Fig. 3).

The difference between Figs. 2 and 3 lies in the following. Draw a straight line inclined at 135° to the x-axis through the point M of intersection of the straight line $y = x$ and the curve $y = \varphi(x)$. This straight line, together with the line $y = x$, will divide the plane into four quadrants. If the curve in the neighbourhood of the point M lies in the left and the right quadrants of the plane and if the initial approximation is taken in this neighbourhood, then the iteration process converges. If, on the other hand, the curve lies in the upper and the lower quadrants of the plane, the process will be divergent.

However, to use this rule one has first to sketch the graph of the function $y = \varphi(x)$, but this is not always expedient. So another convergence test has to be devised for the iteration process which would enable the convergence (or divergence) to be established analytically without any geometric constructions. This test will be discussed in Sec. 10. But first we should become acquainted with the concept of the contraction mapping.

9. Contraction Mappings (Contractions)

Consider the function $y = \varphi(x)$ defined on the interval $[a, b]$. So for every point x_0 of this interval, there is a corresponding point y_0 on the y-axis, namely the point $y_0 = \varphi(x_0)$. To plot this point one has to draw a vertical line through the point x_0 of the x-axis until it intersects the

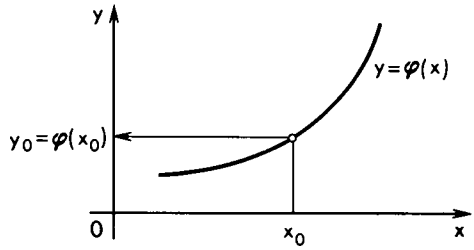


Fig. 4

graph of the function $y = \varphi(x)$ and then to draw a horizontal line through the point of intersection until it intersects the y-axis (Fig. 4). Thus, the function $y = \varphi(x)$ gives a mapping or map of the interval $[a, b]$ to the y-axis. The set of all points on the y-axis corresponding to points of interval $[a, b]$ is called the *image* of the interval. For instance, the image of the interval $[2, 5]$ under the mapping $y = x^2$ is the interval $[4, 25]$, and the image of the interval $[-1, 6]$ under the same

mapping is the interval $[0, 36]$ (draw the graph of the function $y = x^2$). It may be proved that if the function $y = \varphi(x)$ is continuous on the interval $[a, b]$, then the image of this interval will also be an interval on the y-axis. If the function $y = \varphi(x)$ is also a monotonic increasing function, the image of the interval $[a, b]$ is the interval $[\varphi(a), \varphi(b)]$, while if it is a monotonic decreasing function, the image is the interval $[\varphi(b), \varphi(a)]$ (Fig. 5).

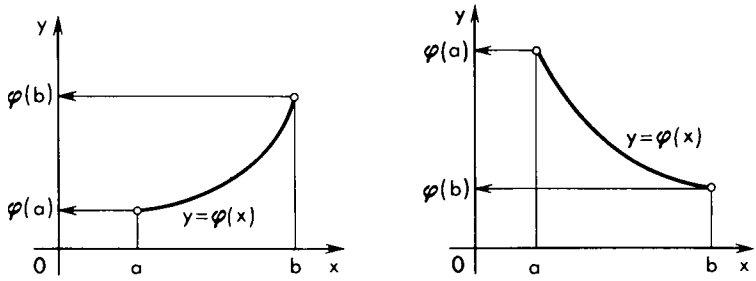


Fig. 5

Instead of considering the mapping of the interval $[a, b]$ to the y-axis one may consider its mapping to the x-axis. To do this, after mapping the interval to the y-axis, rotate the y-axis clock-wise through 90°. As a result, the points of the interval $[a, b]$ will first be mapped to points on the y-axis and then to points on the x-axis. In this way the function $\varphi(x)$ gives a mapping of the interval $[a, b]$ to the x-axis. We shall denote this mapping as follows: $x \to \varphi(x)$. If the function $\varphi(x)$ is continuous, we obtain as a result an interval on the x-axis.

It may happen that the image $[a_1, b_1]$ of the interval $[a, b]$ turns out to be a part of $[a, b]$. For instance, under the mapping $y = \sqrt{x+1}$ the interval $[0, 4]$ maps to a part of this interval, the interval $[1, 3]$. In such cases we shall speak of $\varphi(x)$ mapping the interval $[a, b]$ to a subinterval. If $\varphi(x)$ maps the interval $[a, b]$ to a subinterval $[a_1, b_1]$, then any subinterval $[a, b]$ will map to a subinterval of $[a_1, b_1]$. In particular, the interval $[a_1, b_1]$ will itself be mapped by $\varphi(x)$ to its subinterval $[a_2, b_2]$. In the same way the interval $[a_2, b_2]$ maps to subinterval $[a_3, b_3]$ under the same mapping, and so on. As a result, we obtain a system of intervals:

$$[a, b], [a_1, b_1], \ldots, [a_n, b_n], \ldots$$

each of which is a subinterval of the preceding interval and such that $[a_{n+1}, b_{n+1}]$ is the image of $[a_n, b_n]$ under the mapping $\varphi(x)$.

For instance, the mapping $x \to 1 - \frac{\sqrt{1}}{x+2}$ takes the interval [0,4] to its subinterval [1/2, 5/6]. Applying this mapping to the interval [1/2, 5/6] we obtain the interval [3/5, 11/17], etc. Every successive interval is included in the preceding one.

Two cases are possible: either there is an interval $[c, d]$ common to all intervals $[a_n, b_n]$, or the intervals have only one common point ξ. In the latter case the system of intervals $[a_n, b_n]$ is said to *contract to a point* ξ

Below we shall formulate conditions for the system of intervals $[a, b]$, $[a_1, b_1]$, ..., $[a_n, b_n]$, ... to contract to a point. For this let us introduce the important concept of a contraction. The mapping $\varphi(x)$ which takes interval $[a, b]$ to its subinterval $[a_1, b_1]$ is called a *contraction* if it decreases the distance between any two points of this interval at least M times where $M > 1$. Since the distance between x_2 and x_1 is $|x_2 - x_1|$, the condition may be formulated as follows.

A mapping is a contraction on the interval $[a, b]$ *if there is a number* q, *where* $0 < q < 1$, *such that for any two points* x_1, x_2 *belonging to the interval* $[a, b]$ *the inequality*

$$|\varphi(x_2) - \varphi(x_1)| < q|x_2 - x_1| \tag{32}$$

is satisfied (here $q = 1/M$).

The length of an arbitrary subinterval $[c, d]$ of the interval $[a, b]$ is decreased by a contraction mapping $\varphi(x)$ at least $M = 1/q$ times. Indeed, let $[c_1, d_1]$ be the image of the interval $[c, d]$. Then c_1 and d_1 are the images of some points x_1 and x_2 of the interval $[c, d]$:

$$c_1 = \varphi(x_1), \qquad d_1 = \varphi(x_2)$$

But then

$$|d_1 - c_1| = |\varphi(x_2) - \varphi(x_1)| \leqslant q|x_2 - x_1|$$

Since the points x_1 and x_2 lie in the interval $[c, d]$, the distance between them, $|x_2 - x_1|$, is less than the length $|d - c|$ of the interval $[c, d]$. Therefore

$$|d_1 - c_1| \leqslant q|d - c|$$

Our statement has been proved.

Now we can formulate a condition for the system of intervals $[a_1, b_1]$, ..., $[a_n, b_n]$, ..., obtained from interval $[a, b]$ by successive use of mapping $\varphi(x)$, to contract to a point.

If the mapping $\varphi(x)$ *which takes the interval* $[a, b]$ *to its subinterval* $[a_1, b_1]$ *is a contraction, then the system of intervals* $[a_1, b_1]$, ... $[a_n, b_n]$, ...*will collapse to a point* ξ *belonging to the interval* $[a, b]$.

Indeed, since the mapping $\varphi(x)$ is a contraction, for any n

$$|b_n - a_n| \leqslant q|b_{n-1} - a_{n-1}|$$

In the same way

$$|b_{n-1} - a_{n-1}| \leqslant q|b_{n-2} - a_{n-2}|$$

But then

$$|b_n - a_n| \leqslant q^2|b_{n-2} - a_{n-2}|$$

Repeating this reasoning we obtain

$$|b_n - a_n| \leqslant q^n|b - a|$$

Since $0 < q < 1$, the sequence of numbers $q, q^2, \ldots, q^n, \ldots$ tends to zero and so the lengths $|b_n - a_n|$ of the intervals $[a_n, b_n]$ tend to zero as n tends to infinity. Hence there can be no interval $[c, d]$ which is a subinterval of all the intervals $[a_n, b_n]$. Therefore the system of intervals

$$[a, b], [a_1, b_1], \ldots, [a_n, b_n], \ldots$$

contracts to a point.

Finally, let us consider mappings $\varphi(x)$ for which inequality (32),

$$|\varphi(x_2) - \varphi(x_1)| < q|x_2 - x_1|$$

is satisfied for any pair of numbers x_2 and x_1. Such mappings are contractions on the whole number axis. Let us demonstrate that in this case there is an interval which contracts under the mapping $\varphi(x)$. Since the condition (32) is satisfied for any two points x_1 and x_2, it suffices to show that there is an interval which $\varphi(x)$ maps into itself. Take an arbitrary number a and put $b = \varphi(a)$. Choose the number $q_1 < 1$ so that $q < q_1$.

Let us put

$$R = \frac{|b - a|}{1 - q_1}$$

We shall show that the interval $[a - R, a + R]$ is taken by the mapping $\varphi(x)$ into a subinterval. Indeed, let x be a point of this interval. Then $|x - a| < R$. By virtue of inequality (32) we conclude that

$$|\varphi(x) - b| = |\varphi(x) - \varphi(a)| < q|x - a| \leqslant qR.$$

But then

$$|\varphi(x) - a| = |\varphi(x) - b + b - a| \leqslant |\varphi(x) - b| + |b - a| \leqslant$$
$$\leqslant qR + |b - a| = qR + (1 - q_1)R = (1 + q - q_1)R < R$$

This demonstrates that any point of the interval $[a - R, a + R]$ is taken by the mapping $\varphi(x)$ to a point of the same interval and that, consequently, the mapping $\varphi(x)$ contracts the interval $[a - R, a + R]$.

10. Contraction Mappings and Method of Iteration

Let us now return to the method of iteration. This method is used in solving equations of the type $x = \varphi(x)$. If ξ is a root of this equation, then $\xi = \varphi(\xi)$, and the mapping $x \to \varphi(x)$ leaves the point ξ fixed where it is. Hence, *the problem of solving the equation* $x = \varphi(x)$ *is equivalent to the problem of finding the fixed points of the mapping* $\varphi(x)$.

If the mapping $\varphi(x)$ is a contraction on the interval $[a, b]$, there is always a fixed point in this interval. To convince ourselves of this let us take a set of intervals

$$[a_1, b_1], [a_2, b_2], \ldots, [a_n, b_n], \ldots$$

obtained from $[a, b]$ by successive use of the mapping $\varphi(x)$. Since $\varphi(x)$ is a contracton mapping on the interval $[a, b]$, there is a unique point ξ common to all the intervals $[a_n, b_n]$. This is a fixed point of the mapping $\varphi(x)$.

Indeed, the mapping $\varphi(x)$ takes every interval $[a_n, b_n]$ to a subinterval $[a_{n+1}, b_{n+1}]$. Therefore, the image $\varphi(x)$ of any point x of the interval $[a_n, b_n]$ lies in the subinterval $[a_{n+1}, b_{n+1}]$ and so certainly inside $[a_n, b_n]$. Since the point ξ belongs to all the intervals $[a_n, b_n]$, its image $\varphi(\xi)$ must also belong to all these intervals. But the only point which belongs to all the intervals $[a_n, b_n]$ is the point ξ. Therefore, $\varphi(\xi) = \xi$, i.e. ξ is a fixed point of the mapping $\varphi(x)$.

Thus, for contraction mappings on the interval $[a, b]$ there is always a fixed point lying in the interval. This point is unique. Indeed, should there be another fixed point η, $\eta = \varphi(\eta)$, the inequality

$$|\eta - \xi| = |\varphi(\eta) - \varphi(\xi)| < q|\eta - \xi|$$

would hold. Since $0 < q < 1$, this inequality can be satisfied only if $|\eta - \xi| = 0$, i.e if $\eta = \xi$.

Now we can formulate a necessary condition for the convergence of the iteration process.

Suppose the function $\varphi(x)$ yields a contraction mapping on the interval $[a, b]$. Then for any point x_0 belonging to this interval the sequence of numbers $x_0, x_1, x_2, \ldots, x_n, \ldots$ where $x_{n+1} = \varphi(x_n)$ converges to a root ξ of the equation $x = \varphi(x)$ which lies in this interval.

Indeed, let $[a_n, b_n]$, $n = 1, 2, \ldots$, be a sequence of intervals obtained from $[a, b]$ by sequential applications of the mapping $\varphi(x)$. Since the point x_0 lies in the interval $[a, b]$, its image $x_1 = \varphi(x_0)$ lies in the interval $[a_1, b_1]$, the image $x_2 = \varphi(x_1)$ of the point x_1 lies in the interval $[a_2, b_2]$, and so on. Thus, for any n the point x_n lies in the interval $[a_n, b_n]$. Since the lengths of the intervals $[a_n, b_n]$ approach zero as n increases, the sequence of points $x_1, \ldots, x_n, \ldots$ approaches the common point ξ of these intervals.

The above reasoning shows that any point x_0 of the interval $[a, b]$ may be taken as the initial point.

Let us now find the rate at which the points $x_0, x_1, \ldots, x_n, \ldots$ approach the point ξ. Since $\xi = \varphi(\xi)$, we have for any point c of the interval $[a, b]$

$$|\varphi(c) - \xi| = |\varphi(c) - \varphi(\xi)| < q|c - \xi| \tag{33}$$

Apply inequality (33) to the points $x_0, \ldots, x_n, \ldots$. Since $x_n = \varphi(x_{n-1})$, it follows that

$$|x_n - \xi| = |\varphi(x_{n-1}) - \xi| < q|x_{n-1} - \xi|$$

But then for any n we have

$$|x_n - \xi| < q|x_{n-1} - \xi| < q^2|x_{n-2} - \xi| < \ldots < q^n|x_0 - \xi|$$

Hence, *the error $|x_n - \xi|$ decreases with increasing n at least as fast as a geometric progression having ratio q.*

Let us give some examples of how the condition proved above may be applied.

Example 1. Can the iteration method be applied to the solution of the equation

$$x = \frac{1}{4 + x^2} \tag{34}$$

In this case

$$\varphi(x) = \frac{1}{4 + x^2}$$

For arbitrary x_1 and x_2, we have

$$|\varphi(x_2) - \varphi(x_1)| = \left|\frac{1}{4+x_2^2} - \frac{1}{4+x_1^2}\right| =$$

$$= \frac{|x_1^2 - x_2^2|}{(4+x_2^2)(4+x_1^2)} = \frac{|x_1 + x_2|}{|4+x_1^2||4+x_2^2|}|x_2 - x_1|$$

Using the inequality between the geometric mean and the arithmetic mean we obtain

$$|x| = \frac{1}{2}\sqrt{4x^2} \leqslant \frac{4+x^2}{4}$$

Therefore

$$|x_1 + x_2| \leqslant |x_1| + |x_2| \leqslant \frac{(4+x_1^2)+(4+x_2^2)}{4} =$$

$$= 2 + \frac{x_1^2 + x_2^2}{4} \leqslant 2 + \frac{x_1^2 + x_2^2}{2} + \frac{x_1^2 x_2^2}{8} =$$

$$= \frac{1}{8}(4+x_1^2)(4+x_2^2)$$

We have proved that for any x_1 and x_2 the inequality

$$\frac{x_1 + x_2}{(4+x_1^2)(4+x_2^2)} \leqslant \frac{1}{8}$$

holds and that therefore

$$|\varphi(x_2) - \varphi(x_1)| \leqslant \frac{1}{8}|x_2 - x_1|$$

This means that the mapping $\varphi(x)$ is a contraction on the entire axis.

We know already that in this case there is an interval which is mapped by this contraction into itself. To find it put $a = 0$. The mapping $\varphi(x)$ takes the point $a = 0$ to the point $b = 1/4$. Furthermore, in our case $q = 1/8$. Put $q_1 = 1/4$ and denote the number $\dfrac{|b-a|}{1-q_1} = \dfrac{1}{3}$ by R. The interval $\left[-\dfrac{1}{3}, \dfrac{1}{3}\right]$ is mapped by $\varphi(x)$ into itself. Consequently, there is one fixed point in this interval, which is a root of equation (34). To find this point take an arbitrary point of the interval $\left[-\dfrac{1}{3}, \dfrac{1}{3}\right]$, for instance the point $x_0 = 0$. Using the method of iterations we obtain

$$x_1 = \frac{1}{4} = 0.25$$

$$x_2 = \frac{1}{4 + 0.25^2} = \frac{1}{4.0625} = 0.2461$$

$$x_3 = \frac{1}{4 + 0.2461^2} = \frac{1}{4.0605} = 0.2463$$

$$x_4 = \frac{1}{4 + 0.2463^2} = \frac{1}{4.0605} = 0.2463$$

So within an accuracy of 0.0001 we have $x_3 = x_4$. It follows that, with an accuracy of 0.0001, the root of equation (34) lying in the interval $\left[-\frac{1}{3}, \frac{1}{3}\right]$ is 0.2463. Since the mapping $\varphi(x)$ is a contraction on the whole axis, equation (34) has no other roots.

Example 2. Can the method of successive approximations be used to solve the equation

$$x = 1 + \sqrt[3]{x}$$

in the interval $[-1, 8]$?

Here $\varphi(x) = 1 + \sqrt[3]{x}$. Since $\varphi(-1) = 0$ and $\varphi(8) = 3$, $\varphi(x)$ maps the interval $[-1, 8]$ into itself. However, on this interval it is not a contraction since, for instance, if $x_1 = -0.008$, $x_2 = 0.008$

$$|\varphi(x_2) - \varphi(x_1)| = \left|\sqrt[3]{0.008} - \sqrt[3]{-0.008}\right| = 0.4 > |x_2 - x_1|$$

In proving that the mapping of Example 1 was a contraction we used the inequality $\sqrt{ab} < \frac{a+b}{2}$. Now we shall introduce some inequalities which often have to be used to prove that some mapping is a contraction.

Prove that for $x > 0$ the inequality

$$\sin x < x < \tan x \tag{35}$$

holds. To do this note that the area S_{OAB} (Fig. 6) of the sector OAB with central angle x lies between the areas of the triangles OAB and OAT

$$S_{\Delta OAB} < S_{\text{sect } OAB} < S_{\Delta OAT}$$

But

$$S_{\Delta OAB} = \frac{R^2 \sin x}{2}, \qquad S_{\Delta OAT} = \frac{R^2 \tan x}{2}$$

(R is the radius of the circle) The area of the sector OAB is $\frac{R^2 x}{2}$ (the angle is measured in radians). Therefore

$$\frac{R^2 \sin x}{2} < \frac{R^2 x}{2} < \frac{R^2 \tan x}{2}$$

Cancelling $R^2/2$ we obtain inequality (35). From inequality (35) it follows that for $0 < x < 1$ we have

$$x < \arcsin x$$

and for $x > 0$

$$x > \arctan x$$

Note also the inequalities

$$e^x > 1 + x, \quad x > 0,$$
$$\ln(1 + x) < x, \quad 0 < x < 1$$

which are somewhat more difficult to prove

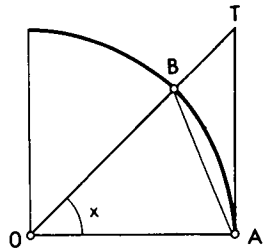


Fig 6

Example 3. Find out whether the equation

$$x = 1 + \frac{1}{2} \arctan x \tag{36}$$

can be solved by the method of iterations.

Since for all values of x we have $1 + \frac{1}{2}\arctan x > 0$, the equation can have only positive roots. We have

$$\varphi(x) = 1 + \frac{1}{2}\arctan x \tag{37}$$

Therefore

$$|\varphi(x_2) - \varphi(x_1)| = \left|\left(1 + \frac{1}{2}\arctan x_2\right) -$$

$$-\left(1 + \frac{1}{2}\arctan x_1\right)\right| = \frac{1}{2}|\arctan x_2 - \arctan x_1|$$

But for $x_1 \geqslant 0,\ x_2 \geqslant 0$

$$\arctan x_2 - \arctan x_1 = \arctan \frac{x_2 - x_1}{1 + x_1 x_2}$$

and therefore

$$|\varphi(x_2) - \varphi(x_1)| = \frac{1}{2}\left|\arctan \frac{x_2 - x_1}{1 + x_1 x_2}\right| <$$

$$< \frac{1}{2}\left|\frac{x_2 - x_1}{1 + x_1 x_2}\right| < \frac{1}{2}|x_2 - x_1|$$

It follows that the mapping is a contraction on the semiaxis $[0, \infty)$. It maps the interval $[0, \sqrt{3}]$ on its subinterval $[1, 1 + \pi/6]$. Therefore there is unique root of equation (36) which lies in the interval $[1, 1 + \pi/6]$. To find this root put $x_1 = 1$. Then

$$x_2 = 1 + \frac{1}{2}\arctan 1 = 1 + \frac{\pi}{8} \approx 1.39$$

$$x_3 = 1 + \frac{1}{2}\arctan 1.39 = 1.474$$

$$x_4 = 1 + \frac{1}{2}\arctan 1.474 = 1.487$$

$$x_5 = 1 + \frac{1}{2}\arctan 1.487 = 1.489$$

$$x_6 = 1 + \frac{1}{2}\arctan 1.489 = 1.490$$

$$x_7 = 1 + \frac{1}{2}\arctan 1.490 = 1.490$$

We see that the equality $x_6 = x_7 = 1.490$ is satisfied with an accuracy of 0.001. This means that with this accuracy the root of our equa-

tion is 1.490. Since the mapping $\varphi(x)$ is a contraction on the entire semiaxis $0 \leqslant x < \infty$, equation (36) has no other roots.

If often happens that an equation $x = \varphi(x)$ which cannot be solved by means of the method of iterations can be transformed to an equation which makes the use of this method possible. Let us take, for instance, the equation

$$x = x^3 - 2 \tag{38}$$

Since here we have

$$\varphi(1) = -1 < 1, \quad \varphi(2) = 6 > 2$$

this equation has a root lying in the interval $[1, 2]$. But the mapping $x^3 - 2$ is not a contraction on this interval since it does not map it into a subinterval. Rewrite equation (38) in the form

$$x = \sqrt[3]{x+2}$$

Putting $\psi(x) = \sqrt[3]{x+2}$ we have

$$|\psi(x_2) - \psi(x_1)| = \left|\sqrt[3]{x_2+2} - \sqrt[3]{x_1+2}\right| =$$

$$= \left|\frac{x_2 - x_1}{\sqrt[3]{(x_2+2)^2} + \sqrt[3]{(x_1+2)(x_2+2)} + \sqrt[3]{(x_1+2)^2}}\right|$$

In the interval $[1, 2]$ we have $x_1 \geqslant 1$, $x_2 \geqslant 1$. Therefore

$$|\psi(x_2) - \psi(x_1)| \leqslant \frac{1}{3\sqrt[3]{9}}|x_2 - x_1|$$

We have proved that the mapping $\psi(x)$ is a contraction on the interval $[1, 2]$. Put $x_1 = 1$ and apply the method of iterations.

$$x_2 = \sqrt[3]{3} = 1.442$$
$$x_3 = \sqrt[3]{3.442} = 1.510$$
$$x_4 = \sqrt[3]{3.510} = 1.520$$
$$x_5 = \sqrt[3]{3.520} = 1.521$$
$$x_6 = \sqrt[3]{3.521} = 1.521$$

Hence, with an accuracy of 0.001, 1.521 is the root of equation (38) lying inside the interval $[1, 2]$. The equation has no other roots.

We see that an appropriate transformation of the initial equation has reduced it to the form to which the method of iterations is applicable.

The convergence test for the method of iterations described herein is not very convenient to use since it requires proof of rather complex inequalities. Below (Sec. 21) we shall consider one corollary of this test which makes the proof of convergence of the iteration process much easier

11. Method of Chords

The method of iterations is one of the most general methods for the approximate solution of equations. Many other methods of approximate solution of equations are just particular cases of the method of iterations. Now we shall describe one of these methods, called the *method of chords* (rule of false positions).

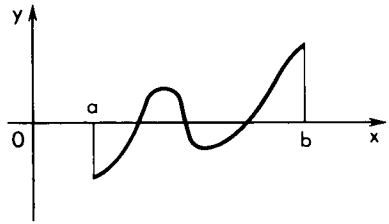


Fig. 7

Suppose we have to solve the equation $f(x)=0$. This problem is equivalent to that of finding points in which the graph of the function $y=f(x)$ intersects the x-axis. Suppose the function $f(x)$ is continuous and its values at points a and b have different signs. Then for at least one point of interval $[a, b]$ the function vanishes. In other words, the graph $y=f(x)$ intersects the x-axis in at least one point ξ of the interval $[a, b]$. In general, there may be several such points (Fig. 7). However, if the function $y=f(x)$ is *monotonic* on the interval $[a, b]$ and its values at the ends of the interval have an opposite sign, then the graph of this function intersects the x-axis in only one point ξ. To find this point approximately, substitute the chord MN for the arc of the curve $y=f(x)$ on the interval $[a, b]$, and find the point of intersection T of this chord with the x-axis (Fig. 8).

To do this consider similar triangles MM_1T and NN_1T. From the similarity of these triangles it follows that $\frac{M_1T}{MM_1}=\frac{TN_1}{N_1N}$. But from

Fig. 8 it may be seen that $M_1T = a_1 - a$, $TN_1 = b - a_1$, $MM_1 = -f(a)$ and $N_1N = f(b)$, where a_1 denotes the abscissa of the point of intersection of the chord MN with the x-axis. Therefore

$$\frac{a_1 - a}{-f(a)} = \frac{b - a_1}{f(b)}$$

Solving this equation we obtain

$$a_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}$$

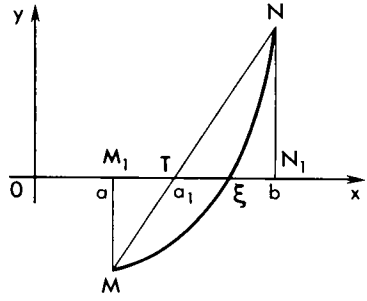


Fig. 8

This may also be written in the form:

$$a_1 = b - f(b)\frac{b - a}{f(b) - f(a)} \tag{39}$$

or

$$a_1 = a - f(a)\frac{b - a}{f(b) - f(a)} \tag{40}$$

[check by reducing the right-hand sides of formulas (39) and (40) to a common denominator].

The number a_1 is the approximate value of the root of the equation $f(x) = 0$ lying between the points a and b.

Since the signs of the numbers $f(a)$ and $f(b)$ are opposite, there are two possibilities: either the sign of $f(a)$, or the sign of $f(b)$ is opposite to that of $f(a_1)$. If the signs of the function $f(x)$ at points a and a_1 are opposite, formula (39) is applied to the interval $[a, a_1]$ and the following approximation is obtained for the root sought:

$$a_2 = a_1 - f(a_1)\frac{a_1 - a}{f(a_1) - f(a)} \tag{41}$$

If, on the other hand, the function $f(x)$ assumes values having opposite signs at points a_1 and b, then formula (40) is applied to the interval $[a_1, b]$ by taking

$$a_2 = a_1 - f(a_1)\frac{b - a_1}{f(b) - f(a_1)} \tag{42}$$

After finding the value of a_2, formula (39) is applied to the interval $[a, a_2]$ (or formula (40) is applied to the interval $[a_2, b]$ as the case may be) and the next approximation a_3 is found. Generally, if the approximation a_n has already been found, the next approximation is found from the formula

$$a_{n+1} = a_n - f(a_n)\frac{a_n - a}{f(a_n) - f(a)} \tag{43}$$

or the formula

$$a_{n+1} = a_n - f(a_n)\frac{b - a_n}{f(b) - f(a_n)} \tag{44}$$

We have obtained two formulae (43) and (44). Let us see now when each one should be used. Suppose the curve is concave up. In this case the points of the curve should be joined to whichever of its ends M or N, where the function is positive. If, on the other hand, the curve is concave down, the points should be joined to the end where the function is negative. The different situations which might arise are illustrated in Fig. 9. These drawings make the following statement geometrically obvious:

Suppose the function $f(x)$ is continuous and monotonic on the interval $[a, b]$, the direction of its concavity is constant and it assumes values of opposite sign at the ends of the interval. Then, provided the right approximation formula has been chosen, the method of chords gives a sequence of points converging to the root of the equation $f(x) = 0$.

If, on the other hand, the choice of the formula is made incorrectly the method of chords may take the point a_2 outside the interval $[a, b]$. This is illustrated by Fig. 10.

The method of chords just described is a particular case of the method of iterations. Suppose the function $f(x)$ does not become zero at $x = a$. In this case the equation $f(x) = 0$ is equivalent to the equation

$$x = x - f(x)\frac{x - a}{f(x) - f(a)} \tag{45}$$

Indeed, if $f(\xi)=0$ then

$$\xi=\xi-f(\xi)\frac{\xi-a}{f(\xi)-f(a)} \tag{46}$$

Conversely, if $\xi\neq a$ and if equation (46) holds, then $f(\xi)=0$.

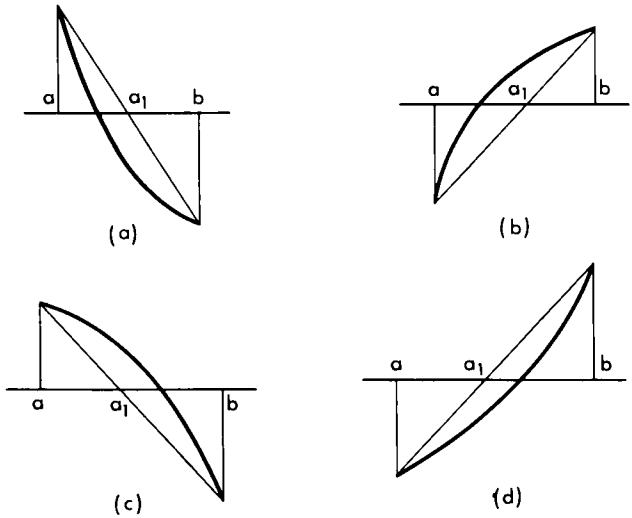


Fig 9

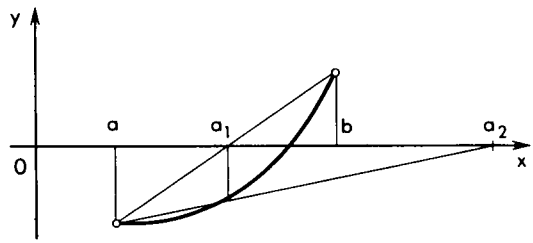


Fig 10

But equation (45) is of the form $x=\varphi(x)$ where

$$\varphi(x)=x-f(x)\frac{x-a}{f(x)-f(a)}=\frac{af(x)-xf(a)}{f(x)-f(a)}$$

Put $x_0=b$ and apply the method of iterations. We shall obtain the same sequence of numbers $a_1, a_2, \ldots, a_n, \ldots$ as is obtained in using the

method of chords:

$$a_{n+1} = a_n - f(a_n)\frac{a_n - a}{f(a_n) - f(a)}$$

As an example let us solve, using the method of chords, the equation

$$x^3 + 3x - 1 = 0 \tag{47}$$

Here $f(x) = x^3 + 3x - 1$. Since $f(0) = -1$, $f(1) = 3$, equation (47) has at least one root in the interval $[0, 1]$. If we plot the graph of the function $y = x^3 + 3x - 1$ we can see that on the interval $[0, 1]$ it is concave up Therefore we use formula (39). According to this formula the first approximation for this root is the number

$$x_1 = b - f(b)\frac{b - a}{f(b) - f(a)} = 1 - 3\frac{1 - 0}{3 - (-1)} = 0\,25$$

To find the second approximation we use the formula

$$x_2 = b - f(b)\frac{b - x_1}{f(b) - f(x_1)} = 1 - 3\frac{1 - 0.25}{3 + 0\,23} = 0.31$$

Then

$$x_3 = 1 - 3\frac{1 - 0.31}{3 + 0.040} = 0.319$$

$$x_4 = 1 - 3\frac{1 - 0.319}{3 + 0.010} = 0.322$$

$$x_5 = 1 - 3\frac{1 - 0.322}{3 + 0.0006} = 0.322$$

Hence, with an accuracy of 0 001 the root of the equation in the interval $[0, 1]$ is 0.322.

12. Improved Method of Chords

If the method of chords converges, its rate of convergence is the same as that of the method of iterations — the error in the value of the root decreases as a geometric progression. There is a way of improving the method of chords so that its rate of convergence becomes much greater. In the usual method of chords we use at each step one of

the ends of the interval $[a, b]$ and the last approximation obtained Instead the last two approximations may be used, for they are closer to the root sought than the ends of the interval $[a, b]$.

The formula which makes use of the last two approximations is of the following form (Fig. 11*a*)·

$$a_{n+1} = a_n - f(a_n)\frac{a_n - a_{n-1}}{f(a_n) - f(a_{n-1})} \tag{48}$$

Here a_1 is computed with the aid of formula (39) and a_2 — with the aid of formula (41) or (42), depending on the signs of $f(a)$, $f(b)$ and $f(a_1)$ if

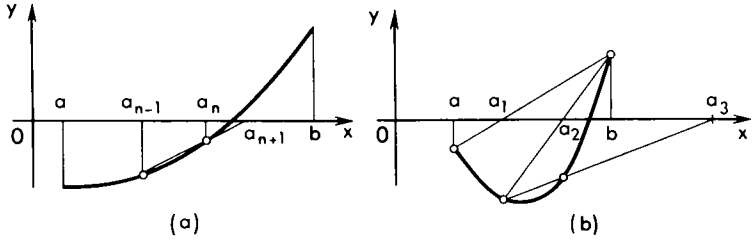


Fig 11

$f(a) < 0$, $f(b) > 0$, then for $f(a_1) < 0$ formula (42) is chosen while for $f(a_1) > 0$ formula (41) is chosen.

If by chance it turns out that point a_3 computed with the aid of formula (48) lies outside the interval $[a, b]$, during the next step the end of the interval closest to it should be taken instead of this point (Fig. 11*b*).

The convergence of the improved method of chords turns out to be much better than that of the usual method, i e. if ξ is the root of the equation $f(x) = 0$, then

$$|a_{n+1} - \xi| < C|a_n - \xi|^t \tag{49}$$

where

$$t = \frac{1 + \sqrt{5}}{2} \approx 1\,618$$

As an example let us use this method to solve the same equation

$$x^3 + 3x - 1 = 0$$

which we solved above using the method of chords The first approxi-

mations $a_1 = 0.25$ and $a_2 = 0.31$ are the same as in the usual mothod of chords.

The next approximation is computed with the aid of the formula

$$a_3 = a_2 - f(a_2)\frac{a_2 - a_1}{f(a_2) - f(a_1)} =$$

$$= 0\,31 + 0.040\frac{0\,31 - 0.25}{-0.040 + 0.234} = 0\,3223$$

We have $f(0\,3223) = 0\,0004$. Clearly, $x = 0.3223$ is the root sought to an accuracy of 0.0001.

13. Derivative of Polynomial

In solving the equation $f(x) = 0$ with the aid of the method of iterations much depends on how the equation is reduced to the form $x = \varphi(x)$. In many cases the best method is that proposed by Newton. It is based on the concept of the derivative. In this section we shall talk about what is meant by the derivative of a polynomial. This will enable us to use Newton's method for the solution of algebraic equations, i.e. of equations of the form

$$a_0x^k + a_1x^{k-1} + \ \dots \ + a_k = 0 \tag{50}$$

Let

$$f(x) = a_0x^k + a_1x^{k-1} + \ \dots \ + a_k$$

be a polynomial Consider the polynomial $f(x + \alpha)$, i.e. the expression

$$a_0(x + \alpha)^k + a_1(x + \alpha)^{k-1} + \ \dots \ + a_k \tag{51}$$

If we remove the brackets in expression (51) we find that in some of the terms there are no α's at all, some terms contain them raised to the first power, some to the second, etc. Let us group together the terms containing α to the same power. Then the polynomial $f(x + \alpha)$ will assume the form

$$f(x + \alpha) = f_0(x) + f_1(x)\alpha + f_2(x)\alpha^2 + \ \dots \ + f_k(x)\alpha^k \tag{52}$$

(since the degree of the polynomial $f(x)$ is k the highest power of α in the expansion (52) is also k). Evidently, $f_0(x), \dots, f_k(x)$ are polynomials in x as well.

Example. Let

$$f(x) = 2x^3 - 3x^2 + 6x - 1$$

Then

$$f(x+\alpha) = 2(x+\alpha)^3 - 3(x+\alpha)^2 + 6(x+\alpha) - 1 =$$
$$= 2(x^3 + 3x^2\alpha + 3x\alpha^2 + \alpha^3) - 3(x^2 + 2x\alpha + \alpha^2) +$$
$$+ 6(x+\alpha) - 1 = (2x^3 - 3x^2 + 6x - 1) +$$
$$+ (6x^2 - 6x + 6)\alpha + (6x - 3)\alpha^2 + 2\alpha^3$$

Consequently, in this case

$$f_0(x) = 2x^3 - 3x^2 + 6x - 1$$

$$f_1(x) = 6x^2 - 6x + 6$$

$$f_2(x) = 6x - 3$$

$$f_3(x) = 2$$

We see that the term $f_0(x)$ coincides with $f(x)$. This is not a chance coincidence. If we put $\alpha = 0$ in equation (52) we obtain $f(x) = f_0(x)$.

Let us now turn to the next term $f_1(x)\alpha$. The coefficient of α, i.e. the polynomial $f_1(x)$, is called the *derivative of the polynomial* $f(x)$. For instance, the derivative of the polynomial $2x^3 - 3x^2 + 6x - 1$ is $6x^2 - 6x + 6$. The derivative of a polynomial is usually written as $f'(x)$.

Hence, *the derivative* $f'(x)$ *of the polynomial* $f(x)$ *is the coefficient of* α *in the expansion of the polynomial* $f(x+\alpha)$ *in powers of* α.

Using the notation introduced above we may re-write formula (52) in the following form:

$$f(x+\alpha) = f(x) + f'(x)\alpha + \ldots \tag{53}$$

The dots in the above formula denote terms containing $\alpha^2, \alpha^3, \ldots, \alpha^k$. For instance,

$$2(x+\alpha)^3 - 3(x+\alpha)^2 + 6(x+\alpha) - 1 = 2x^3 -$$
$$- 3x^2 + 6x - 1 + (6x^2 - 6x + 6)\alpha + \ldots$$

We have introduced the concept of the derivative of a polynomial $f(x)$. Let us now show how to calculate this derivative. To do this consider the polynomial

$$f(x+\alpha) = a_0(x+\alpha)^k + a_1(x+\alpha)^{k-1} + \ldots$$
$$\ldots + a_{k-1}(x+\alpha) + a_k$$

Substituting for each term the expression $(x+\alpha)^m = x^m + mx^{m-1}\alpha + \ldots$ (see Sec. 6) we obtain

$$f(x+\alpha) = a_0(x^k + kx^{k-1}\alpha + \ldots) +$$
$$+ a_1[x^{k-1} + (k-1)x^{k-2}\alpha^2 + \ldots] + \ldots$$
$$\ldots + a_{k-1}(x+\alpha) + a_k = a_0x^k + a_1x^{k-1} + \ldots$$
$$\ldots + a_k + \alpha[ka_0x^{k-1} + (k-1)a_1x^{k-2} + \ldots + a_{k-1}] + \ldots$$

Comparing this equation with (53),

$$f(x+\alpha) = f(x) + \alpha f'(x) + \ldots$$

we can make the following statement:

The derivative of a polynomial

$$f(x) = a_0x^k + a_1x^{k-1} + \ldots + a_{k-1}x + a_k \tag{54}$$

is of the form

$$f'(x) = ka_0x^{k-1} + (k-1)a_1x^{k-2} + \ldots + a_{k-1} \tag{55}$$

For instance, the derivative of the polynomial

$$f(x) = 6x^7 + 8x^3 - 3x^2 - 1$$

is

$$f'(x) = 42x^6 + 24x^2 - 6x$$

14. Newton's Method for Approximate Solution of Algebraic Equations

Let us return now to the approximate solution of algebraic equations. Suppose we have an equation

$$a_0x^k + a_1x^{k-1} + \ldots + a_k = 0 \tag{56}$$

Suppose we have somehow managed to find an approximate value x_1 of the root of this equation. We shall show how a more accurate value of this root can be found. Let α_1 be the error in the value x_1, i.e. suppose $x_1 + \alpha_1$ is the root of equation (56). Then we must have

$$a_0(x_1+\alpha_1)^k + a_1(x_1+\alpha_1)^{k-1} + \ldots + a_k = 0 \tag{57}$$

In other words,

$$f(x_1+\alpha_1) = 0$$

where $f(x)$ denotes the polynomial

$$a_0x^k + a_1x^{k-1} + \ \ldots \ + a_k$$

But according to formula (31) we have

$$f(x_1 + \alpha_1) = f(x_1) + \alpha_1 f'(x_1) + \ \ldots$$

where the dots denote terms containing $\alpha_1^2, \ldots, \alpha_1^k$. Hence, to determine α_1 we have the equation

$$f(x_1 + \alpha_1) = f(x_1) + \alpha_1 f'(x_1) + \ \ldots = 0 \tag{58}$$

If the initial approximation x_1 is good enough, its error α_1 will be small. In this case the terms in equation (58) denoted by dots will be small in comparison with α_1. Neglecting these terms we obtain an approximate equation to determine α_1:

$$f(x_1) + \alpha_1 f'(x_1) \approx 0 \tag{59}$$

It follows from this that

$$\alpha_1 \approx -\frac{f(x_1)}{f'(x_1)} \tag{60}$$

Therefore the formula for the improved value x_2 of the root of our equation will be

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \tag{61}$$

Then we may again improve the approximation obtained. The formula for the third approximation is

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

In general, if the n-th approximation x_n for the root sought has been found, the formula for the next approximation is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{62}$$

In detailed form the formula is written down as follows:

$$x_{n+1} = x_n - \frac{a_0x_n^k + a_1x_n^{k-1} + \ldots + a_{k-1}x_n + a_k}{ka_0x_n^{k-1} + (k-1)a_1x_n^{k-2} + \ldots + a_{k-1}} \tag{63}$$

If the approximate values of x_n and x_{n+1} coincide within the accuracy specified, our process (within the limits of accuracy specified) will be concluded and the value of the root sought will have been found.

The method of solving equations described above is due to the famous English mathematician Newton.

Newton's method is closely connected with the method of iterations. Specifically, if the functions $y = f(x)$ and $y = f'(x)$ have no common roots, the equation $f(x) = 0$ will be equivalent to the equation

$$x = x - \frac{f(x)}{f'(x)} \tag{64}$$

Applying the method of iterations to this equation we obtain a sequence of numbers $x_1, x_2, \ldots, x_n$ connected by the same relation

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{65}$$

as in Newton's method. In other words, Newton's method consists in writing down the equation $f(x) = 0$ in the form of (64) and applying to it the method of iterations.

Example. Use Newton's method to solve the equation

$$x^3 - 3x - 5 = 0$$

with an accuracy of 0.001 choosing the first approximation $x_1 = 3$.

Since the derivative of the polynomial

$$f(x) = x^3 - 3x - 5$$

is the polynomial

$$f'(x) = 3x^2 - 3$$

formula (62) assumes the form

$$x_{n+1} = x_n - \frac{x_n^3 - 3x_n - 5}{3x_n^2 - 3}$$

Therefore

$$x_2 = 3 - \frac{27 - 9 - 5}{27 - 3} = 3 - \frac{13}{24} = 2.46$$

$$x_3 = 2.46 - \frac{14.89 - 7.38 - 5}{18.16 - 3} = 2.46 - 0.165 = 2.295$$

$$x_4 = 2.295 - \frac{12.088 - 6.885 - 5}{15.801 - 3} = 2.295 - 0.016 = 2.279$$

$$x_5 = 2\,279 - \frac{11.837 - 6\,807 - 5}{15\,582 - 3} = 2\,279$$

We see that with an accuracy of 0.001

$$x_4 = x_5$$

Therefore the root of the equation $x^3 - 3x - 5 = 0$ is equal to 2.279 with an accuracy of 0.001.

The method of approximate computation of roots presented in Sec. 6 is a particular case of Newton's method. Indeed, finding $\sqrt[k]{a}$ just means solving the equation

$$x^k - a = 0$$

But the derivative of the polynomial $x^k - a$ is kx^{k-1} and therefore formula (62) for the equation

$$x^k - a = 0$$

takes the form

$$x_{n+1} = x_n - \frac{x_n^k - a}{kx_n^{k-1}} = \frac{a + (k-1)x_n^k}{kx_n^{k-1}}$$

This is just the formula that was used to compute the approximations for $\sqrt[k]{a}$.

Note the following essential difference between the solution of the equation $x^k - a = 0$ and the solution of the general algebraic equation

$$a_0x^k + a_1x^{k-1} + \ \ldots \ + a_k = 0$$

For the equation $x^k - a = 0$ the choice of the initial approximation x_1 was of no importance. No matter what value of x_1 was chosen, after a certain number of steps we obtained the value of $\sqrt[k]{a}$ with the accuracy specified. In the case of the solution of equation (56) the situation is different. Here one initial value results in one root, another initial value results in a different root while some initial values result in no definite value at all—the sequence of numbers $x_1, x_2, \ldots, x_n, \ldots$ computed using formula (62) does not tend to any definite limit, that is, it diverges.

15. Geometrical Meaning of Derivative

So far we have given Newton's method only for algebraic equations. In order to generalize it to apply to equations of arbitrary form we shall have to generalize the concept of the derivative and introduce

it for functions of all types. To do this let us explain the geometrical meaning of the derivative.

Consider the graph of the polynomial

$$y = a_0x^k + a_1x^{k-1} + \ \ldots \ + a_k$$

and take two points M and N on this graph (Fig. 12). Let the abscissa of the point M be x and the abscissa of the point N be $x + \alpha$. Then the

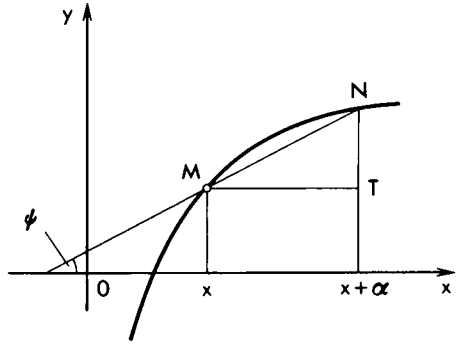


Fig. 12

ordinates of the points M and N will be given by the expressions

$$f(x) = a_0x^k + a_1x^{k-1} + \ \ldots \ + a_k$$

and

$$f(x + \alpha) = a_0(x + \alpha)^k + a_1(x + \alpha)^{k-1} + \ \ldots \ + a_k$$

respectively. Draw a secant through the points M and N and calculate its slope k_{sec}*. It may be seen from the drawing that

$$\tan\psi = \frac{TN}{MT}$$

But the segment MT is equal to the difference in the abscissas of the points M and N and therefore

$$MT = (x + \alpha) - x = \alpha$$

The segment TN is equal to the difference in the ordinates of these

* By the *slope of a line* is meant the tangent of the angle of inclination of the line with the positive direction of the x-axis. For instance, if a line makes an angle of 60° with the x-axis, its slope is equal to $\sqrt{3}$.

points and therefore

$$TN = f(x + \alpha) - f(x)$$

It follows that

$$\tan \psi = \frac{TN}{MT} = \frac{f(x + \alpha) - f(x)}{\alpha}$$

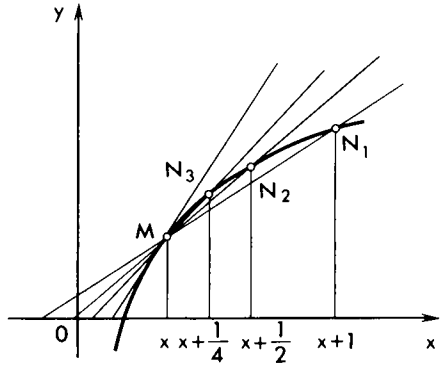


Fig. 13

But by formula (53)

$$f(x + \alpha) = f(x) + \alpha f'(x) + \ \ldots$$

where the dots denote terms containing α^2, α^3, ... Therefore

$$\tan \psi = \frac{\alpha f'(x) + }{\alpha} = f'(x) + \ .$$

where the dots now denote terms containing α, α^2, . .

Hence, the slope of the secant MN is expressed by the formula

$$k_{\text{sec}} = \tan \psi = f'(x) + \ \ldots \tag{66}$$

Let us now start decreasing the value of α. In doing this secant MN will turn around the point M. In the limiting case, when $\alpha = 0$, the secant will become the tangent to the curve $y = f(x)$ at the point M. Figure 13 shows the positions of the secant for $\alpha = 1$; 1/2; 1/4.

But when $\alpha = 0$ all the terms denoted by dots in formula (66) vanish. Therefore the slope of the tangent to the graph of a polynomial $y = f(x)$ at the point with abscissa x is expressed by the formula

$$k_{\tan} = f'(x) \tag{67}$$

Hence, *the derivative of a polynomial $f(x)$ is equal to the slope of the tangent to the graph of the polynomial at the point with abscissa x.*

Example. Find the angle that the tangent to the graph of the polynomial

$$f(x) = x^3 - 4x^2 + 5x + 1$$

drawn at the point $x = 2$, makes with the x-axis.

Since

$$f(x) = 3x^2 - 8x + 5$$

$f'(2) = 1$. Consequently, $\tan \varphi = 1$ and the angle is $\varphi = 45°$.

16. Geometrical Meaning of Newton's Method

We can now clarify the geometrical meaning of Newton's method for the approximate solution of algebraic equations. Suppose we need to solve the equation $f(x) = 0$ where $f(x)$ is a polynomial. Geometrically, this problem is the problem of finding the points of intersection of the graph of the function $y = f(x)$ with the x-axis, i.e. the points where $y = 0$

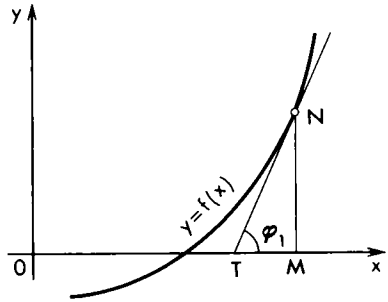


Fig. 14

Suppose an approximate value of the root of this equation x_1 has already been found. Draw a tangent to the curve $y = f(x)$ at the point N with abscissa x_1. If the choice of x_1 is fortunate, the point T of intersection of the tangent with the x-axis will be nearer to the point of intersection of the curve $y = f(x)$ with the x-axis than the point M (Fig. 14).

To find the abscissa x_2 of point T consider the triangle TMN. The side MN of the right-angled triangle is just the value of the function

$y = f(x)$ at the point x_1, i.e. $MN = f(x_1)$. On the other hand, the side TM is equal to $x_1 - x_2$. Hence, the tangent of the angle φ_1 which the tangent line makes with the x-axis is expressed by the formula

$$\tan \varphi_1 = \frac{f(x_1)}{x_1 - x_2} \tag{68}$$

From (68) it follows that

$$x_2 = x_1 - \frac{f(x_1)}{\tan \varphi_1} \tag{69}$$

But $\tan \varphi_1$ is the slope of the tangent to the curve $y = f(x)$ drawn at the point having abscissa x_1. Therefore, in accordance with the geometrical meaning of the derivative, $\tan \varphi_1 = f'(x_1)$.

Hence, formula (69) may be re-written in this form:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Thus we have found the second approximation for the root sought. Let us now draw a tangent to the curve $y = f(x)$ at the point with abscissa x_2. The abscissa of the point of intersection of this tangent with the x-axis is given by the formula

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

In general, if the approximation x_n has already been found, then in order to obtain the next approximation x_{n+1} one should draw a tangent to the curve $y = f(x)$ at the point with abscissa x_n. The abscissa of the point of intersection of this tangent with the x-axis will then provide us with the value of x_{n+1}.

The formula for calculating x_{n+1} is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{70}$$

or, equivalently,

$$x_{n+1} = x_n - \frac{f(x_n)}{\tan \varphi_n} \tag{70'}$$

where φ_n is the angle that the tangent to the curve $y = f(x)$ at the point with abscissa x_n makes with the x-axis. This formula coincides with

formula (62) of Newton's method. Thus, we have found the geometrical meaning of Newton's method. The essence of it is the substitution of the tangent to the curve $y = f(x)$ for its arc. For this reason another name for Newton's method is the *method of tangents.*

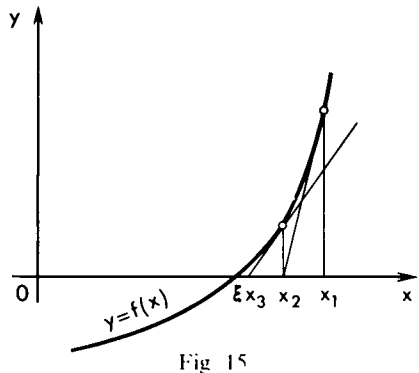


Fig. 15

Figure 15 shows how the points $x_1, x_2, \ldots, x_n$, obtained using Newton's method, approach the point ξ of intersection of the curve $y = f(x)$ with the x-axis.

17. Derivatives of Arbitrary Functions

The geometrical interpretation of Newton's method given above enables us to generalize it to apply to any equations of the form $f(x) = 0$, where $f(x)$ may now be some function other than a polynomial. To find the solution of this equation take some approximate value x_1 of its root. Draw a tangent to the curve $y = f(x)$ at the point with abscissa x_1 and denote its point of intersection with the x-axis by x_2. Draw a new tangent to the curve $y = f(x)$ at the point having abscissa x_2, etc. It may easily be established that, as in the case of a polynomial,

$$x_{n+1} = x_n - \frac{f(x_n)}{\tan \varphi_n} \tag{71}$$

where $\tan \varphi_n$ is the slope of the tangent to the curve $y = f(x)$ at the point with abscissa x_n.

Formula (71) is still of no use for calculations because we don't know how to find $\tan \varphi_n$. So we have to learn to calculate the slope of tangents drawn to graphs of arbitrary functions $y = f(x)$ (and not only

to graphs of polynomials). Let us first find the slope of the secant. Let M be a point on the graph of the function $y=f(x)$ and MN a secant passing through this point. Applying the same reasoning as in the case of polynomials we conclude that the slope of the secant is given by the formula

$$k_{\text{sec}}=\tan\psi=\frac{f(x+\alpha)-f(x)}{\alpha} \tag{72}$$

where x is the abscissa of the point M and $x+\alpha$ the abscissa of the point N If we decrease α, the secant will turn around the point M until it eventually occupies the position of the tangent to the curve $y=f(x)$ at this point (see Fig. 12). Therefore we may write that

$$k_{\tan}=\tan\varphi=\lim_{\alpha\to 0}\frac{f(x+\alpha)-f(x)}{\alpha} \tag{73}$$

We shall call the limit on the right-hand side the *derivative* of the function $f(x)$, and denote it by $f'(x)$, i e we shall put

$$f'(x)=\lim_{\alpha\to 0}\frac{f(x+\alpha)-f(x)}{\alpha} \tag{74}$$

Now we may write the equality (73) in the form

$$k_{\tan}=\tan\varphi=f'(x) \tag{75}$$

Thus, the derivative $f'(x)$ of any function at some point (not just of a polynomial) is equal to the slope of the tangent drawn at this point to the curve $y=f(x)$*.

Since $\tan\varphi_n=f'(x_n)$ formula (71) may be re-written in the form

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \tag{76}$$

This formula coincides with formula (62). Thus, Newton's method has been extended to apply to all equations of the form $f(x)=0$.

*If it is impossible to draw a tangent to the graph of the function $y=f(x)$ at a point with abscissa x (for instance, if the graph has a break at this point, i.e. it bends sharply at an angle) then the function has no derivative at this point

18. Computation of Derivatives

We have seen in the preceding section that to find the slope of the tangent to the curve $y=f(x)$ we have to compute the limit

$$f'(x)=\lim_{\alpha\to 0}\frac{f(x+\alpha)-f(x)}{\alpha}$$

This computation is, generally speaking, quite difficult. But for many important cases the limit has already been worked out. In other words, the derivatives of the more frequently used functions are known. Below is a list of the most common derivatives.

1 $(a)'=0$

2. $(x^k)'=kx^{k-1}$

3 $(a^x)'=a^x\ln a$

4. $(\sin ax)'=a\cos ax$

5. $(\cos ax)'=-a\sin ax$

6. $(\tan ax)'=\dfrac{a}{\cos^2 ax}$

7. $(\cot ax)'=-\dfrac{a}{\sin^2 ax}$

8 $(\log_a x)'=\dfrac{1}{x\ln a}$

9 $(\arcsin ax)'=\dfrac{a}{\sqrt{1-a^2x^2}}$

10. $(\arctan ax)'=\dfrac{a}{1+a^2x^2}$

($\ln a$ in formulas 3 and 8 denotes the logarithm to the base $e=2.71828\ldots$—the so-called *natural*, or *Napierian logarithm*). Note that k in formula 2 may be not only a natural number, but any real number. For example,

$$(\sqrt{x})'=(x^{1/2})'=\frac{1}{2}x^{1/2-1}=\frac{1}{2\sqrt{x}}$$

$$\left(\frac{1}{x^2}\right)'=(x^{-2})'=-2x^{-3}=-\frac{2}{x^3}$$

The formulas 1-10 are insufficient to calculate the derivatives of all functions. However, if the function $f(x)$ is made up by using arithmetical operations of functions whose derivatives we know how to find then we may easily find its derivative. To do this we use the following rules which are proved (together with formulas 1-10) in higher mathematics courses

1. *The derivative of the sum of two functions is equal to the sum of their derivatives, i.e.*

$$[f_1(x)+f_2(x)]'=f_1'(x)+f_2'(x)$$

2. *A constant multiplier may be taken outside the derivative sign:*

$$[af(x)]' = af'(x)$$

3. *The formula for calculating the derivative of a product of two functions is*

$$[f_1(x)f_2(x)]' = f_1'(x)f_2(x) + f_1(x)f_2'(x)$$

4. *The formula for calculating the derivative of a fraction is*

$$\left[\frac{f_1(x)}{f_2(x)}\right]' = \frac{f_1'(x)f_2(x) - f_1(x)f_2'(x)}{[f_2(x)]^2}$$

The rule given in Sec. 13 for calculating the derivative of a polynomial is a corollary of rules 1 and 2 and formula 2 of the list.

Example 1. Find the derivative of the fraction

$$f(x) = \frac{3x^2 - x + 1}{2x^3 + 5}$$

Applying rule 4 we obtain

$$f'(x) = \frac{(3x^2 - x + 1)'(2x^3 + 5) - (3x^2 - x + 1)(2x^3 + 5)'}{(2x^3 + 5)^2}$$

Next, applying the rule for the differentiation of a polynomial, we find

$$(3x^2 - x + 1)' = 6x - 1$$

and

$$(2x^3 + 5)' = 6x^2$$

and therefore

$$f'(x) = \frac{(2x^3 + 5)(6x - 1) - (3x^2 - x + 1)6x^2}{(2x^3 + 5)^2} =$$

$$= \frac{-6x^4 + 4x^3 + 6x^2 + 30x - 5}{(2x^3 + 5)^2}$$

Example 2. Find the derivative of the function

$$f(x) = \frac{1}{10}\left(\arcsin 3x - \frac{1}{x^2}\right)$$

Solution. Using formulas 2 and 9 and rules 1 and 2 we obtain

$$f'(x) = \frac{1}{10}\frac{3}{\sqrt{1 - 9x^2}} - \frac{1}{10}\left(\frac{-2}{x^3}\right) = \frac{3}{10\sqrt{1 - 9x^2}} + \frac{1}{5x^3}$$

Example 3. Find the derivative of the function

$$f(x) = 10^x \sin 2x$$

Applying rule 3 and formulas 3 and 4 we obtain

$$f'(x) = (10^x)' \sin 2x + 10^x (\sin 2x)' =$$
$$= 10^x \sin 2x \ln 10 + 10^x \cdot 2 \cos 2x =$$
$$= 10^x (\sin 2x \ln 10 + 2 \cos 2x)$$

The rules presented above enable the derivative to be found in many cases. There is one other very important rule—the rule for the calculation of the derivative of a complicated function. It is formulated as follows:

If a function $y = f(x)$ may be written in the form $y = F(z)$ where $z = \varphi(x)$, then its derivative is given by the formula

$$f'(x) = F'(z)\varphi'(x) \tag{77}$$

where $z = \varphi(x)$.

Example. Find the derivative of the function $y = \sin(x^3)$. This function may be written down in the form $y = \sin z$ where $z = x^3$. The derivative of the function $F(z) = \sin z$ is $F'(z) = \cos z$ and the derivative of the function $\varphi(x) = x^3$ is $\varphi'(x) = 3x^2$. Using formula (77) we obtain

$$[\sin(x^3)]' = F'(z)\varphi'(x) = \cos z \cdot 3x^2$$

Substituting for z its value $z = x^3$ we obtain

$$[\sin(x^3)]' = 3x^2 \cos(x^3)$$

The reader can find a more detailed explanation of the concept of the derivative, for instance, in the book by Ya. B. Zeldovich *Higher Mathematics for the Beginner.*

19. Finding the First Approximations

Let us deal now with the choice of initial approximations. When solving the equation $f(x) = 0$ the first approximation may be found graphically. For this one should plot the function $y = f(x)$ and find the points of intersection of the graph with the x-axis (at these points $y = 0$ and therefore $f(x) = 0$).

If for some reason or other it is inconvenient to plot the function (for instance, if the equation is being solved by a computer), another method of finding the first approximation is employed In this method

the values of the function for some values of its argument are computed (for instance, for integral values of the argument lying inside definite bounds). If the function $y = f(x)$ is continuous (i e. its graph has no discontinuities), then there is a root of the equation $f(x) = 0$ lying between the values a and b of the argument for which the function assumes values of opposite sign (Fig. 16*a*). If the graph of the function contains discontinuities, it may happen that the function moves from the negative values to positive ones in a jump without passing through zero on the way (Fig. 16*b*) The values a or b may be taken as first approximations for the root of the equation $f(x) = 0$

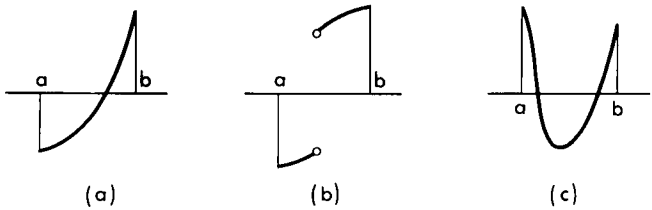


Fig 16

Note that by this method we may leave out some of the roots of the equation. For instance, Fig. 16*c* depicts the case when the function $y = f(x)$ has the same sign at two points, but vanishes between them.

We have thus obtained two points: a and b. To find out which of them should best be chosen as the initial approximation x in Newton's method, consider Fig. 17. Figures 17*a* and 17*b* show that *if the curve is concave up whichever of the two points a and b for which the function* $f(x)$ *is positive should be chosen as the initial approximation. A different choice of the initial approximation may even result in the point* x_2 *being outside the interval* $[a, b]$. Similarly, *if the curve is concave down, the point where the function* $f(x)$ *is negative* (Fig. 17*c, d*) *should be chosen as the initial approximation.*

This rule may be conveniently used if the graph of the function $y = f(x)$ is known. If, on the other hand, the function has not been plotted, then additional calculations are necessary to determine the direction of the function's concavity. In order to do this, the second derivative of the function $f(x)$ should be found. By the *second derivative of the function* $f(x)$ is meant the derivative of its first derivative.

For instance, if we are given the function

$$f(x) = x^3 - 4x^2 + 3x - 1$$

its first derivative is

$$f'(x) = 3x^2 - 8x + 3$$

and the second

$$f''(x) = 6x - 8$$

Higher mathematics courses contain the proof that if the second derivative is positive, the curve on this interval is concave up. If, on the

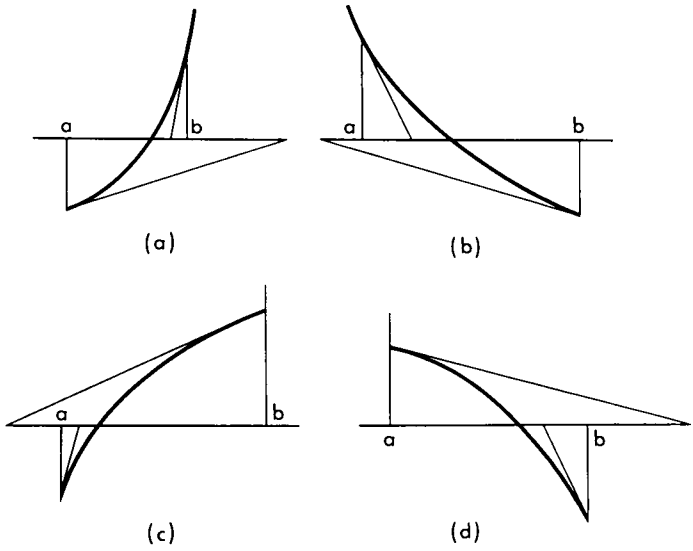


Fig 17

other hand, the second derivative in the interval $[a, b]$ is negative, the curve is concave down. Making use of this information we obtain the following rule for the use of Newton's method:

Let the function $f(x)$ have opposite signs at points a and b and let the second derivative of the function $f(x)$ be positive on the interval $[a, b]$. Then for the initial approximation x_1 one should choose that of the points a or b for which the function $f(x)$ assumes a positive value. If, on the other hand, the second derivative is negative on the interval $[a, b]$, then the point where the function $f(x)$ assumes a negative value should be chosen for the initial approximation x_1.

20. Combined Method of Solving Equations

In solving equations Newton's method is often combined with the method of chords. If the graph of the function $y = f(x)$ is concave up, then the points a_1 and x_1 are found using formulas

$$a_1 = a - f(a)\frac{b-a}{f(b)-f(a)} \tag{78}$$

$$x_1 = b - \frac{f(b)}{f'(b)} \tag{79}$$

If, on the other hand, the graph of the function $y = f(x)$ is concave down, then formula (78) is used to find the point a_1 and the formula

$$x_1 = a - \frac{f(a)}{f'(a)} \tag{80}$$

to find x_1.

As may be seen from Fig. 18*a* and 18*b*, the root ξ of the equation $f(x) = 0$ usually lies between the points a_1 and x_1. After Newton's

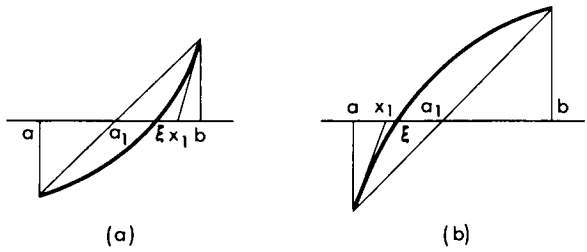


Fig 18

method and the method of chords are again applied, we get a new pair of points a_2 and x_2, etc.

In this way two sequences of points are obtained: $a_1, a_2, \ldots a_n, \ldots$ and $x_1, x_2, \ldots, x_n, \ldots$ which approach the root ξ sought from different sides. The advantage of the method described is that it approximates the roots from both above and below.

Example. Using the combined method solve the equation

$$x - \sin x - 0.5 = 0$$

with an accuracy of 0.001.

Compile a table of values of the continuous function

$$f(x) = x - \sin x - 0.5$$

x	-1	0	1	2
$f(x)$	$-0\,659$	-0.5	-0.341	0.591

It follows from this table that the root of this equation lies between 1 and 2. Using formulas 2 and 4 of Sec. 18 we obtain

$$f'(x) = 1 - \cos x$$

Therefore Newton's formula assumes in our case the form

$$x_{n+1} = x_n - \frac{x_n - \sin x_n - 0.5}{1 - \cos x_n} \tag{81}$$

To find out which of the values, 1 or 2, should be taken for x_0 let us find the second derivative of the function $f(x)$. According to formula 5 of Sec. 18, it is of the form $f''(x) = \sin x$. But on the interval $[1, 2]$ the function $\sin x$ is positive*. Consequently, using the rule established above, the value 2 for which the function $f(x)$ is positive should be taken for x_0.

According to formula (81) we have

$$x_1 = 2 - \frac{2 - \sin 2 - 0.5}{1 - \cos 2} = 2 - \frac{2 - 0\,909 - 0\,5}{1 + 0.416} = 1.583$$

On the other hand, using formula (78) we obtain

$$a_1 = 1 - (-0.341)\frac{2 - 1}{0\,591 - (-0.341)} = 1.366$$

Then applying formulas (81) and (78) to the interval $[a_1, x_1]$ we obtain

$$x_2 = 1\,583 - \frac{1.583 - 1\,000 - 0\,5}{1 + 0.012} = 1.501$$

*The function $\sin x$ is positive on the interval $[0, \pi] = [0, 3.141\ \ldots]$. Therefore $\sin x$ is positive also on the subinterval $[1, 2]$ of this interval.

and

$$a_2 = 1\,366 + 0.113\,\frac{1\,583 - 1\,366}{0.083 + 0.113} = 1.491$$

Next we find

$$x_3 = 1.498$$
$$a_3 = 1.498$$

Hence the root of our equation with an accuracy of 0.001 is 1.498.

21. Convergence Test for Method of Iterations

Let us now apply the concept of the derivative to the derivation of a new convergence test for the method of iterations. We shall need for this purpose another formula called the Lagrange formula (Lagrange was a French mathematician of the 18th century)

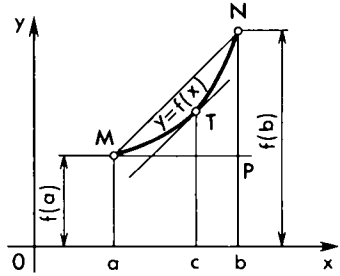


Fig. 19

Consider the curve $y = f(x)$ on the interval $[a, b]$. Denote by M the initial point of this curve and by N its end point and draw a chord MN. The slope of this chord is

$$k_{\text{chord}} = \tan\psi = \frac{PN}{MP}$$

(Fig. 19) But $MP = b - a$ and $PN = f(b) - f(a)$, and so

$$k_{\text{chord}} = \frac{f(b) - f(a)}{b - a}$$

Let us now denote by T the point of the arc MN furthest from the chord MN. If we draw a straight line parallel to the chord through this

point, it will be tangent to the curve, for if it intersected the curve it would mean that there were points more distant from the chord MN than the point T. In other words, the tangent to the curve at point T is parallel to the chord MN and has therefore the same slope as the chord. But the slope of the tangent is equal to $f'(c)$ where c is the abscissa of the point T. So we have the formula

$$f'(c) = \frac{f(b) - f(a)}{b - a} \tag{82}$$

This formula is called the Lagrange formula. Note that the point c in the Lagrange formula always lies between the points a and b. The Lagrange formula may also be written in the form

$$f(b) - f(a) = f'(c)(b - a) \tag{83}$$

Let us return now to the solution of the equation $x = \varphi(x)$ by the method of iterations. Suppose the mapping $y = \varphi(x)$ maps the interval $[a, b]$ into itself so that on this interval the inequality $|\varphi'(x)| < q$ holds where q is a number less than unity, $q < 1$. Take any two points x_1 and x_2 of the interval $[a, b]$. Then the points $\varphi(x_1)$ and $\varphi(x_2)$ will also belong to the interval $[a, b]$. Using the Lagrange formula we obtain

$$\varphi(x_2) - \varphi(x_1) = \varphi'(c)(x_2 - x_1)$$

where c is some point lying between x_1 and x_2 and so belonging to the interval $[a, b]$. The inequality $|\varphi'(c)| < q < 1$ holds and therefore

$$|\varphi(x_2) - \varphi(x_1)| \leqslant q|x_2 - x_1| \tag{84}$$

Inequality (84) shows that $\varphi(x)$ is a contraction mapping. But we know that if $x \to \varphi(x)$ is a contraction mapping of the interval $[a, b]$ into itself, then for any point x_0 of this interval the sequence $x_0, x_1, \ldots, x_n, \ldots$ where $x_{n+1} = \varphi(x_n)$ converges to the root of the equation $x = \varphi(x)$. Thus, we have proved the following theorem.

Theorem. *Let the function $y = \varphi(x)$ be the mapping of the interval $[a, b]$ into itself and suppose in this interval the inequality $|\varphi'(x)| < q$, where $q < 1$, holds. Then for any point x_0 of the interval $[a, b]$ the sequence of points $x_0, x_1, \ldots, x_n, \ldots$, where $x_{n+1} = \varphi(x_n)$, converges to the root of the equation $x = \varphi(x)$.*

Roughly speaking the theorem just proved says that the process of successive approximations enables us to find those roots ξ of the equation $x = \varphi(x)$ for which the inequality $|\varphi'(\xi)| < 1$ is satisfied. One could say that these points attract the broken line (open polygon) which geometrically depicts the iteration process (see Sec. 8) while the points for which $|\varphi'(\xi)| > 1$ repulse this line.

If the inequality $|\varphi'(x)| < q < 1$ is satisfied on the entire number axis, the iteration process converges independently of the choice of the initial approximation x_0 (see Sec. 10).

Example 1. Can the iteration process be applied to the solution of the equation

$$x = \frac{\cos x + \sin x}{4}$$

In this case

$$\varphi(x) = \frac{\cos x + \sin x}{4}$$

Therefore

$$\varphi'(x) = \frac{-\sin x + \cos x}{4}$$

But $|\sin x| \leqslant 1$, $|\cos x| \leqslant 1$, therefore

$$|\varphi'(x)| = \left|\frac{-\sin x + \cos x}{4}\right| \leqslant \frac{|\sin x| + |\cos x|}{4} < \frac{1}{2}$$

and the iteration process can be applied.

Example 2. Can the iteration process be applied to solve the equation

$$x = 4 - 2^x \tag{85}$$

The root to be found lies in the interval $[1, 2]$ since the continuous function $y = x - 4 + 2^x$ changes sign in this interval:

$$1 - 4 + 2^1 < 0 \text{ while } 2 - 4 + 2^2 > 0$$

But in this case we have

$$\varphi'(x) = -2^x \ln 2$$

Let us evaluate the expression $2^x \ln 2$ in the interval $[1, 2]$. If

$$1 \leqslant x \leqslant 2, \text{ then } 2 \leqslant 2^x \leqslant 4$$

and therefore

$$2 \ln 2 \leqslant 2^x \ln 2 \leqslant 4 \ln 2$$

From the tables of Napierian logarithms (their base is the number $e \approx$ $\approx 2.78...$) we find $\ln 2 = 0.69 \ldots$. Consequently on the interval $[1, 2]$ the inequality

$$1.38 \ldots \leqslant 2^x \ln 2 \leqslant 2.76 \ldots$$

holds and so the iteration process is inapplicable.

In order to be able to apply the iteration process we can transform equation (85). Rewrite it in the form

$$2^x = 4 - x$$

and take the logarithms to the base 2 of both sides. Then we shall obtain

$$x = \log_2(4 - x)$$

In this case

$$\varphi'(x) = -\frac{1}{(4-x)\ln 2}$$

and on the interval [1, 2] the inequality

$$|\varphi'(x)| < \frac{1}{2\ln 2} = \frac{1}{1.38} < 1$$

holds. (The reader may easily derive the inequality himself.)

Therefore when the equation is written in this way the iteration process converges.

22. Rate of Convergence of Iteration Process*

Let us now use the derivative of the function $\varphi(x)$ to find the rate of convergence of the iteration process for the solution of the equation $x = \varphi(x)$. We want to find the rate of decrease in the errors $\alpha_n = \xi - x_n$ of the approximate values $x_1, \ldots, x_n, \ldots$ of the root ξ.

Note that the equalities $\xi = \varphi(\xi)$ and $x_{n+1} = \varphi(x_n)$ are valid. It follows from them that

$$\alpha_{n+1} = \xi - x_{n+1} = \varphi(\xi) - \varphi(x_n)$$

But using the Lagrange formula we have

$$\varphi(\xi) - \varphi(x_n) = \varphi'(c_n)(\xi - x_n) = \varphi'(c_n)\alpha_n$$

where c_n is a point lying between the points x_n and ξ. Therefore

$$\alpha_{n+1} = \varphi'(c_n)\alpha_n \tag{86}$$

The following conclusion may be drawn from equality (86):

Let ξ be the root of the equation $x = \varphi(x)$ lying in the interval $[a, b]$. If in this interval the inequality $|\varphi'(x)| < q < 1$ is satisfied, and the initial

*The section may be omitted on first reading

approximation x_1 is also chosen in the interval $[a, b]$ then for any n the relation

$$|\alpha_{n+1}| < q^n |\alpha_1| \tag{87}$$

holds.

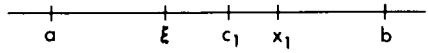


Fig. 20

Indeed, it follows from equality (86) that

$$|\alpha_2| = |\varphi'(c_1)|\,|\alpha_1|$$

But the point c_1 lies in the interval $[a, b]$ (Fig. 20) and therefore

$$|\varphi'(c_1)| < q$$

It follows from this that

$$|\alpha_2| < q|\alpha_1|$$

In the same way we obtain

$$|\alpha_3| = |\varphi'(c_2)|\,|\alpha_2| < q\,|\alpha_2| < q^2\,|\alpha_1|$$

and in general

$$|\alpha_{n+1}| < q^n |\alpha_1|$$

This proves our statement.

Since for $0 < q < 1$ the sequence of numbers $q, q^2, \ldots, q^n \ldots$ tends to zero, the error α_{n+1} also tends to zero as n increases. In other words, with the above assumptions the numbers $x_1, \ldots, x_n, \ldots$ approach the number ξ with the difference $|\xi - x_{n+1}|$ decreasing at a higher rate than $|\alpha_1|q^n$.

In the same way it can be proved that if in the interval $[a, b]$ the inequality

$$|\varphi'(x)| > 1$$

holds the iteration process diverges.

The convergence rate of the iteration process is highest if the derivative of the function $\varphi'(x)$, vanishes at the point ξ. In this case as we get close to ξ, $\varphi'(x)$ tends to zero. Since

$$|\alpha_{n+1}| = |\varphi'(c_n)|\,|\alpha_n|$$

the convergence rate increases as the point ξ is approached.

We have already met with a similar situation when using the method of iteration to extract square roots. Remember that we then substituted the equation $x = \dfrac{x^2 + a}{2x}$ for the equation $x^2 = a$. But the derivative of the function $\varphi(x) = \dfrac{x^2 + a}{2x}$ is

$$\varphi'(x) = \frac{(x^2 + a)'2x - (x^2 + a)(2x)'}{4x^2} = \frac{2x\ 2x - (x^2 + a)2}{4x^2} = \frac{x^2 - a}{2x^2}$$

[see rule 4 of Sec. 18 and formula (55) of Sec. 13], therefore

$$\varphi'(\sqrt{a}) = \frac{(\sqrt{a})^2 - a}{2(\sqrt{a})^2} = 0$$

Hence, the derivative of the function $\varphi(x)$ vanishes at the point $x = \sqrt{a}$ and this accelerates the convergence of the process as the point $x = \sqrt{a}$ is approached.

This increase in the rate of the process as the root of the equation is approached is also a feature of Newton's method (a specific case of which is the method of extracting square roots mentioned above). Indeed, we have already noted that Newton's method involves replacing the equation $f(x) = 0$ by the equation

$$x = x - \frac{f(x)}{f'(x)}$$

and the subsequent solution of this equation by successive approximations. In this case we have

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

but

$$\varphi'(x) = 1 - \left[\frac{f(x)}{f'(x)}\right]' = 1 - \frac{f'(x)[f(x)]' - f(x)[f'(x)]'}{[f'(x)]^2} =$$

$$= 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)''(x)}{[f'(x)]^2}$$

Since at point ξ the equation $f(\xi) = 0$ is satisfied, $\varphi'(\xi) = 0$ and this, as

was shown above, accelerates the convergence of the approximation process as the point ξ is approached.

23. Solving of Systems of Linear Equations by Method of Successive Approximations

Up to now we have been solving equations in one unknown. Now we shall consider the solution of systems of equations starting with systems of equations of the first degree.

Suppose we have m first-degree equations with m unknowns*:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ \dots\ + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ \dots\ + a_{2m}x_m &= b_2 \\ \dots\dots\dots\

We are required to find the unknowns x_1, x_2, x_3 with an accuracy of 0.01.

Let us express the first equation in terms of x_1, the second in terms of x_2, and the third in terms of x_3. Then the system will assume the form

$$\left.\begin{aligned} x_1 &= 0.9 + 0.2x_2 - 0.1x_3 \\ x_2 &= 1.6 - 0.2x_1 + 0.2x_3 \\ x_3 &= 4 - 0.5x_1 - 0.25x_2 \end{aligned}\right\} \tag{89}$$

Take arbitrary values of x_1, x_2, x_3 as the initial approximations, for instance, put $x_1^{(0)} = 0$, $x_2^{(0)} = 0$, $x_3^{(0)} = 0$. Substitute, these values into the right-hand sides of equation (89) and take the values obtained to be the next approximate values of x_1, x_2, x_3. We obtain

$$\begin{aligned} x_1^{(1)} &= 0.9 \\ x_2^{(1)} &= 1.6 \\ x_3^{(1)} &= 4 \end{aligned}$$

Substitute the values thus obtained again into the right-hand sides of the equations (89). We obtain the approximations

$$\begin{aligned} x_1^{(2)} &= 0.9 + 0.2 \times 1.6 - 0.1 \times 4 = 0.82 \\ x_2^{(2)} &= 1.6 - 0.2 \times 0.9 + 0.2 \times 4 = 2.22 \\ x_3^{(2)} &= 4 - 0.5 \times 0.9 - 0.25 \times 1.6 = 3.15 \end{aligned}$$

In general if the values $x_1^{(n)}$, $x_2^{(n)}$, $x_3^{(n)}$ have been found then to obtain the next approximations one has to use the formulas

$$\left.\begin{aligned} x_1^{(n+1)} &= 0.9 + 0.2x_2^{(n)} - 0.1x_3^{(n)} \\ x_2^{(n+1)} &= 1.6 - 0.2x_1^{(n)} + 0.2x_3^{(n)} \\ x_3^{(n+1)} &= 4 - 0.5x_1^{(n)} - 0.25x_2^{(n)} \end{aligned}\right\} \tag{90}$$

The results of the computations are presented in Table 2.

Table 2

n	1	2	3	4	5	6
$x_1^{(n)}$	0.9	0 82	1 03	1 01	1 00	1.00
$x_2^{(n)}$	1 6	2.22	2 07	2 00	1.99	2 00
$x_3^{(n)}$	4.0	3 15	3 03	2 97	3 00	3 00

We see that, with the accuracy required, the following equalities hold:

$$x_1^{(5)} = x_1^{(6)}, \quad x_2^{(5)} = x_2^{(6)}, \quad x_3^{(5)} = x_3^{(6)} \tag{91}$$

Putting in equations (90) $n = 5$ and taking account of the equalities (91) we obtain that within the accuracy reguired

$$x_1^{(5)} \approx 0.9 + 0.2x_2^{(5)} - 0.1x_3^{(5)}$$
$$x_2^{(5)} \approx 1.6 - 0.2x_1^{(5)} + 0.2x_3^{(5)}$$
$$x_3^{(5)} \approx 4 - 0.5x_1^{(5)} - 0.25x_2^{(5)}$$

(in fact, the equations are satisfied exactly but this is unimportant). It follows that the numbers $x_1^{(5)} = 1.00$, $x_2^{(5)} = 2.00$, $x_3^{(5)} = 3.00$ are (within the accuracy required) the solutions of the given system of equations.

The same method is used in the general case*. Suppose we have a system of equations (88). Transpose x_1 in the first equation, x_2 — in the second, etc. Then system (88) assumes the form

$$x_1 = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \ldots - \frac{a_{1m}}{a_{11}}x_m$$
$$x_2 = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \ldots - \frac{a_{2m}}{a_{22}}x_m$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_m = \frac{b_m}{a_{mm}} - \frac{a_{m1}}{a_{mm}}x_1 - \frac{a_{m2}}{a_{mm}}x_2 - \ldots - \frac{a_{m,m-1}}{a_{mm}}x_{m-1} \tag{92}$$

*The text below down to the end of the section may be omitted on first reading.

Let $x_1^{(1)}, \ldots, x_m^{(1)}$ be initial approximations for the unknowns $x_1, \ldots, x_m$

Substituting these values into the right-hand sides of equations (92) we obtain second approximations $x_1^{(2)}, \ldots, x_m^{(2)}$ for the unknowns sought:

$$
\begin{aligned}
x_1^{(2)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} x_2^{(1)} - \ldots - \frac{a_{1m}}{a_{11}} x_m^{(1)} \\
x_2^{(2)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_1^{(1)} - \ldots - \frac{a_{2m}}{a_{22}} x_m^{(1)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_m^{(2)} &= \frac{b_m}{a_{mm}} - \frac{a_{m1}}{a_{mm}} x_1^{(1)} - \ldots - \frac{a_{m,m-1}}{a_{mm}} x_{m-1}^{(1)}
\end{aligned}
$$

In the same way, if the n-th approximations $x_1^{(n)}, x_2^{(n)}, \ldots, x_m^{(n)}$ have already been found, the formulas for the next step are

$$
\begin{aligned}
x_1^{(n+1)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} x_2^{(n)} - \ldots - \frac{a_{1m}}{a_{11}} x_m^{(n)} \\
x_2^{(n+1)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_1^{(n)} - \ldots - \frac{a_{2m}}{a_{22}} x_m^{(n)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_m^{(n+1)} &= \frac{b_m}{a_{mm}} - \frac{a_{m1}}{a_{mm}} x_1^{(n)} - \ldots - \frac{a_{m,m-1}}{a_{mm}} x_{m-1}^{(n)}
\end{aligned}
\qquad (93)
$$

Consider now a problem which deals not with the solution of a system of linear equations by the method of successive approximations, but with finding the limiting state of some process of approximations by means of solving a system of equations.

There are three buckets. The first contains 12 *litres of water, the second and the third are empty. Half of the water in the first bucket is poured into the second, then half of the water in the second bucket is poured into the third, then half of the water contained in the third bucket is poured into the first. This cycle is repeated* 20 *times. It is required to find (with an accuracy of* 0.0001 *l) how much water there will be in each bucket.*

Evidently, this problem deals with successive approximations to some limiting distribution of water. This limiting distribution is characterized by the fact that it does not change as the result of one of the

pouring cycles described. If at the start of a cycle there are x litres of water in the first bucket, y in the second and $12 - x - y$ litres in the third (the total amount of water does not change as the result of the pouring) then the above cycle will be described by the following table:

	1 st bucket	2 nd bucket	3 rd bucket
Initial state	x	y	$12 - x - y$
After 1st pouring	$\frac{x}{2}$	$\frac{x}{2} + y$	$12 - x - y$
After 2nd pouring	$\frac{x}{2}$	$\frac{x}{4} + \frac{y}{2}$	$12 - \frac{3}{4}x - \frac{y}{2}$
After 3rd pouring	$6 + \frac{x}{8} - \frac{y}{4}$	$\frac{x}{4} + \frac{y}{2}$	$6 - \frac{3}{8}x - \frac{y}{4}$

For the amount of water in each bucket to remain unchanged as the result of such a pouring cycle, the following equations should be satisfied:

$$x = 6 + \frac{x}{8} - \frac{y}{4}$$

$$y = \frac{x}{4} + \frac{y}{2}$$

Solving these equations we find that $x = 6$, $y = 3$. Hence, the limiting distribution is such that the first bucket contains 6 l of water, the second — 3 l and, consequently, the third, too, contains 3 l of water

Let us see what the rate is at which a specific distribution of water approaches the limiting values. Suppose the first bucket contains a litres and the second b litres. After one cycle there will be

$$a_1 = 6 + \frac{a}{8} - \frac{b}{4} \tag{94}$$

litres in the first bucket and

$$b_1 = \frac{a}{4} + \frac{b}{2} \tag{95}$$

in the second.

Denote $a - 6$ by α, $b - 3$ by β, $a_1 - 6$ by α_1 and $b_1 - 3$ by β_1. Then it follows from equations (94) and (95) that

$$\alpha_1 = a_1 - 6 = \frac{a-6}{8} - \frac{b-3}{4} = \frac{\alpha}{8} - \frac{\beta}{4}$$

and

$$\beta_1 = b_1 - 3 = \frac{a-6}{4} + \frac{b-3}{2} = \frac{\alpha}{4} + \frac{\beta}{2}$$

After the second cycle the errors will be expressed by the formulas

$$\alpha_2 = \frac{\alpha_1}{8} - \frac{\beta_1}{4} = \frac{1}{8}\left(\frac{\alpha}{8} - \frac{\beta}{4}\right) - \frac{1}{4}\left(\frac{\alpha}{4} + \frac{\beta}{2}\right) = -\frac{3}{64}\alpha - \frac{5}{32}\beta$$

$$\beta_2 = \frac{\alpha_1}{4} + \frac{\beta_1}{2} = \frac{1}{4}\left(\frac{\alpha}{8} - \frac{\beta}{4}\right) + \frac{1}{2}\left(\frac{\alpha}{4} + \frac{\beta}{2}\right) = \frac{5}{32}\alpha - \frac{3}{16}\beta$$

So if $|\alpha| < \varepsilon$ and $|\beta| < \varepsilon$, then

$$|\alpha_2| < \frac{13}{64}\varepsilon \approx 0.2\,\varepsilon$$

$$|\beta_2| < \frac{11}{32}\varepsilon \approx 0\,34\,\varepsilon$$

This means that two pouring cycles decrease the errors α and β at least three times. Therefore after 20 cycles they will decrease at least $3^{10} \approx$ $\approx$ 70 000 times. So with an accuracy of 0.0001 l after 20 pouring cycles there will be 6 l of water in the first bucket, 3 l in the second and 3 l in the third.

24. Solving Systems of Non-linear Equations Using Method of Successive Approximations

The method of successive approximations (iteration) may be used to solve some systems of non-linear equations as well. Consider, for instance, the system

$$\left.\begin{aligned} x &= 2 + \frac{x^2 + y}{20} \\ y &= 1 + \frac{x + y^2}{20} \end{aligned}\right\} \qquad (96)$$

Choose as the first approximations $x_0 = 0$ and $y_0 = 0$. Substituting these approximations for x and y into the right-hand sides of the equations we obtain the following approximations: $x_1 = 2$, $y_1 = 1$. Substi-

tuting these approximations into the right-hand sides of equations (96) we obtain

$$x_2 = 2 + \frac{2^2 + 1}{20} = 2\,25$$

$$y_2 = 1 + \frac{2 + 1^2}{20} = 1\,15$$

Continuing the process we obtain

$$x_3 = 2 + \frac{2\,25^2 + 1\,15}{20} = 2\,31$$

$$y_3 = 1 + \frac{2\,25 + 1\,15^2}{20} = 1\,18$$

$$x_4 = 2 + \frac{2.31^2 + 1\,18}{20} = 2.33$$

$$y_4 = 1 + \frac{2\,31 + 1\,18^2}{20} = 1\,18$$

$$x_5 = 2 + \frac{2\,33^2 + 1\,18}{20} = 2.33$$

$$y_5 = 1 + \frac{2\,33 + 1.18^2}{20} = 1\,18$$

We see that with an accuracy of 0.01 the equalities $x_4 = x_5 = 2.33$ and $y_4 = y_5 = 1.18$ are satisfied. Therefore we have with the accuracy stated above $x = 2.33$ and $y = 1.18$.

In general, if we have a system of equations

$$\left.\begin{aligned} x &= \varphi(x,\ y) \\ y &= \psi(x,\ y) \end{aligned}\right\} \tag{97}$$

where $\varphi(x, y)$ and $\psi(x, y)$ are certain functions, we choose initial approximations x_0 and y_0, substitute them into the right-hand sides of equations (97) and then proceed with the calculations according to the formulas

$$\left.\begin{aligned} x_{n+1} &= \varphi(x_n,\ y_n) \\ y_{n+1} &= \psi(x_n,\ y_n). \end{aligned}\right\} \tag{98}$$

If for some number n the equalities $x_{n+1} \approx x_n$ and $y_{n+1} \approx y_n$ are satisfied with specified accuracy, then we have with that accuracy $x \approx x_n$, $y \approx y_n$.

Systems of equations containing three or more unknowns are solved in the same way.

Let us now establish the conditions guaranteeing the convergence of the process of successive approximations in the solution of systems.

We assume the functions $\varphi(x, y)$ and $\psi(x, y)$ in the system of equations (97) to be defined for some bounded closed region D of the x, y plane. In other words, we will assume that the region D lies entirely inside some square, and that it contains all its boundary points. A circle, a polygon (together with the broken line bounding it), an ellipse, etc. may serve as examples of such regions. We shall, in addition, assume the functions $\varphi(x, y)$ and $\psi(x, y)$ to be continuous in the region D. The functions $\varphi(x, y)$ and $\psi(x, y)$ define a mapping of the region D into some other region lying in the same plane. To find the

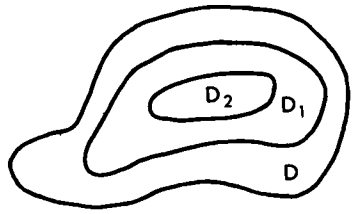


Fig. 21

point into which the point $M_0(x_0, y_0)$ of region D is mapped, one must substitute its coordinates for x and y in $\varphi(x, y)$ and $\psi(x, y)$. The result of such a substitution will be the coordinates of the image of the point M_0. For instance, if

$$\varphi(x, y) = x^2 + y^2$$
$$\psi(x, y) = 2xy$$

then the point $M_0(1, 3)$ will be mapped to the point $N_0(10, 6)$.

In future we shall denote the mapping given by the functions $\varphi(x, y)$ and $\psi(x, y)$ by one letter Φ, and denote the image of the point M under the mapping by $\Phi(M)$. The image of the region D as a whole will be denoted by $\Phi(D)$.

Suppose the mapping Φ maps region D into subregion $D_1 = \Phi(D)$. Then the same mapping may be applied to map the region D_1 into

subregion $D_2 = \Phi(D_1)$ which, of course, also lies inside region D. Continuing this process we obtain a, system of regions D, D_1, ..., D_n, ... (Fig. 21) lying one inside the other.

We shall call the mapping Φ a *contraction* if there is a number q, $0 < q < 1$, such that for any two points M_1 and M_2 inside D the inequality

$$r(\Phi(M_1),\ \Phi(M_2)) \leqslant qr(M_1,\ M_2)$$

holds. Here $r(M, N)$ denotes the distance between the points M and N.

As in the case of a single variable the following proposition can be proved:

Suppose the mapping Φ maps the region D into a subregion, and that it is a contraction. Then there will be a unique point N in the region D such that $N = \Phi(N)$. This point belongs to all the regions D_n. The coordinates ξ, η of the point N satisfy the system of equations (97), *i.e.*

$$\xi = \varphi(\xi,\ \eta)$$
$$\eta = \psi(\xi,\ \eta)$$

As in the case of a single variable, the values ξ and η are computed approximately by the method of iterations. If $M_0(x_0, y_0)$ is an arbitrary point of region D and $M_{n+1} = \Phi(M_n)$ [i.e. $x_{n+1} = \varphi(x_n, y_n)$, $y_{n+1} = \psi(x_n, y_n)$] the sequence of points $M_0, M_1, \ldots, M_n, \ldots$ converges to the fixed point N of the mapping.

25. Modified Distance

The requirement that Φ be a contraction mapping is a sufficient condition for the convergence of the iteration process. But this condition is not necessary. The mapping Φ may not be a contraction, but the iteration process may nevertheless converge. For instance, the mapping Φ defined by the functions $\varphi(x, y) = 1 + 2y$, $\psi(x, y) = 3 + \frac{x}{8}$ is not a contraction. If we take $A(8, 0)$ and $B(8, 4)$ we obtain

$$r(A,\ B) = 4,\ \Phi(A) = (1,\ 4),\ \Phi(B) = (9,\ 4)$$

and

$$r(\Phi(A),\ \Phi(B)) = 8 > r(A,\ B)$$

However, no matter what point M_0 we take, the set of points M_0,

$M_1, \ldots, M_n, \ldots$ converges to the point

$$N\left(9\frac{1}{3},\ 4\frac{1}{6}\right)$$

In some cases the convergence of the iteration process can be established if the definition of the distance between the points in a plane is modified. Actually, there may be different definitions of the distance between two points. For a traveller, it would be natural to measure the distance by the time it takes him to get from point A to point B. In the case depicted in Fig. 22*a* the distance between points A and B will be

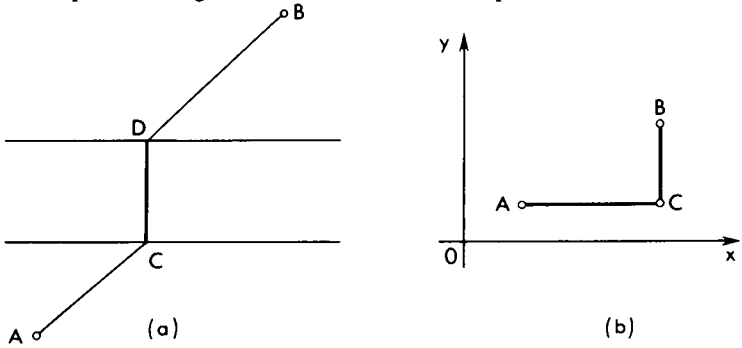


Fig. 22

equal to the sum of the lengths of the segments AC, CD and DB (to get from point A to point B one has to reach the bridge CD, cross it and then walk from point D to point B). If the motion in the plane can take place only along two mutually perpendicular directions, as in Fig. 22*b*, the distance between points A and B should be defined as the sum of the segments AC and CB. In other cases it is more convenient to define the "distance" between points A and B as the length of the greater of the two segments AC and CB. Other definitions of the "distance" between the points of a plane can also be devised. (More detailed information about different definitions of distance can be obtained from the book of Yu. A. Schreider "What is distance".).

The distance between points $r(A, B)$ is usually required to have the following properties:

(1) The distance $r(A, B)$ between any two points A and B is *nonnegative*, being zero only when the points A and B coincide.

(2) For any two points A and B the *symmetry condition* is valid:

$$r(A,\ B) = r(B,\ A)$$

(3) For any three points A, B, C the *triangle inequality* holds:

$$r(A, B) \leqslant r(A, C) + r(C, B)$$

If for some set of objects a distance is defined which has these properties, then the set is called a *metric space*, and the elements of the set are called points of this space. The points of a metric space may even be functions. For continuous functions $\varphi(x)$ and $\psi(x)$ on the interval $[a, b]$ the distance may be defined as the maximum value of the function $|\varphi(x) - \psi(x)|$ on this interval.

$$r(\varphi, \psi) = \max_{a \leqslant x \leqslant b} |\varphi(x) - \psi(x)|$$

We have already seen that there are different ways of turning a plane into a metric space: in all of the above methods of defining distance the conditions (1)-(3) are satisfied.

Here is an interesting example of a case in which the conditions (1) and (3) are satisfied, but the condition of symmetry (2) is not Let us measure the distance between the points A and B of a *mountainous terrain* by the time it takes to get from A to B. Since the time of ascent is not the same as the time of descent, $r(A, B) \neq r(B, A)$

It turns out that the following condition is sufficient for the process of successive approximations to the solution of a system of equations

$$\left.\begin{aligned} x &= \varphi(x, y) \\ y &= \psi(x, y) \end{aligned}\right\} \tag{99}$$

to converge:

The mapping Φ defined by the functions, $\varphi(x, y)$ and $\psi(x, y)$ maps the region D into itself and is a contraction for at least one "distance" $r(A, B)$. In other words, there should be a number q where $0 < q < 1$, such that for any two points M_1 and M_2 of D the inequality

$$r(\Phi(M_1), \Phi(M_2)) \leqslant q r(M_1, M_2)$$

holds.

Take, for instance, the functions $\varphi(x, y) = 1 + 2y$ and $\psi(x, y) = 3 + \frac{x}{8}$. The mapping Φ defined by them turns out to be a contraction with factor $1/2$ if the distance between the points $A(x, y_1)$ and

$B(x_2, y_2)$ is defined by the formula

$$r(A, B) = \left|\frac{1}{2}(x_2 - x_1) + 2(y_2 - y_1)\right| + \left|\frac{1}{2}(x_2 - x_1) - 2(y_2 - y_1)\right|$$

For instance, for points $A(8, 0)$ and $B(8, 4)$ we have $r(A, B) = 16$ while for their images $\Phi(A)$ and $\Phi(B)$

$$r(\Phi(A), \Phi(B)) = 8$$

As a result of the mapping the "distance" becomes twice as small. This is the reason why the process of iterations in the solution of the system of equations

$$x = 1 + 2y$$

$$y = 3 + \frac{x}{8}$$

converges although the mapping Φ is not a contraction with respect to the usual distance (see p. 82)

26. Convergence Tests for Process of Successive Approximations for Systms of Linear Equations

Let us apply the convergence test devised above to systems of linear equations. By choosing different types of "distance" we shall obtain the convergence tests for such systems expressed as properties of their coefficients.

Consider, to begin with, a system of two linear equations in two unknowns

$$\left.\begin{aligned} a_{11}x + a_{12}y &= b_1 \\ a_{21}x + a_{22}y &= b_2 \end{aligned}\right\} \tag{100}$$

Let $a_{11} \neq 0$ and $a_{22} \neq 0$. Solve the first equation for x and the second for y. We obtain a system of equations

$$x = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}y$$

$$y = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x$$

For conciseness put $\frac{b_1}{a_{11}} = \beta_1$, $-\frac{a_{12}}{a_{11}} = \alpha_1$, $\frac{b_2}{a_{12}} = \beta_2$, $-\frac{a_{21}}{a_{22}} = \alpha_2$

Then the system assumes the form

$$\left.\begin{aligned} x &= \alpha_1 y + \beta_1 \\ y &= \alpha_2 x + \beta_2 \end{aligned}\right\} \qquad (100')$$

In this case the functions defining the mapping Φ are of the form

$$\varphi(x, y) = \alpha_1 y + \beta_1, \qquad \psi(x, y) = \alpha_2 x + \beta_2$$

Let us find out what the coefficients α_1 and α_2 should be for this mapping to be a contraction.

As is well known, the distance $r(A, B)$ between points $A(x_1, y_1)$ and $B(x_2, y_2)$ is expressed by the formula

$$r(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The mapping Φ maps the point A to the point $A_1(\alpha_1 y_1 + \beta_1, \alpha_2 x_1 + \beta_2)$ and the point B to the point $B_1(\alpha_1 y_2 + \beta_1, \alpha_2 x_2 + \beta_2)$. The distance between these points is expressed by the formula

$$\begin{aligned} r(A_1, B_1) &= \sqrt{(\alpha_1 y_2 - \alpha_1 y_1)^2 + (\alpha_2 x_2 - a_2 x_1)^2} = \\ &= \sqrt{\alpha_1^2 (y_2 - y_1)^2 + \alpha_2^2 (x_2 - x_1)^2} \end{aligned} \qquad (101)$$

Denote the larger of the numbers $|\alpha_1|$ and $|\alpha_2|$ by q:

$$q = \max(|\alpha_1|, |\alpha_2|)$$

Then the inequality

$$r(A_1, B_1) \leqslant q \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = qr(A, B)$$

follows from formula (101).

Consequently, if $q < 1$, then the mapping Φ is a contraction on the entire plane. In that case, as we know, the process of successive approximations is convergent.

Thus, we have proved that if

$$\max(|\alpha_1|, |\alpha_2|) < 1 \qquad (102)$$

the process of successive approximations for the solution of the system of equations (100') is always convergent.

Let us express the convergence test derived above directly in terms of the coefficients of system of equations (100). To do this recall that

$$\alpha_1 = -\frac{a_{12}}{a_{11}}$$

$$\alpha_2 = -\frac{a_{21}}{a_{22}}$$

Substituting these expressions into the condition (102) we arrive at the following conclusion:

For the process of successive approximations to the solution of a system of linear equations

$$a_{11}x + a_{12}y = b_1$$
$$a_{21}x + a_{22}y = b_2$$

to be convergent it is sufficient that the condition

$$\max\left(\left|\frac{a_{12}}{a_{11}}\right|, \left|\frac{a_{21}}{a_{22}}\right|\right) < 1$$

be satisfied

This condition means that the diagonal coefficients should be greater than the non-diagonal coefficients in the respective lines. For this reason, for instance, when solving the system

$$x - 3y = -11$$
$$6x + y = 10$$

the first equation should be solved for y and the second for x:

$$y = \frac{11}{3} + \frac{1}{3}x, \qquad x = \frac{5}{3} - \frac{y}{6}$$

Sometimes it is useful to carry out a preliminary transformation of the system of equations, substituting for the unknowns x and y other unknowns proportional to them. Take, for instance, the system of equations

$$\left.\begin{aligned} 12x + y &= 14 \\ 3x - 2y &= -1 \end{aligned}\right\} \qquad (103)$$

For it

$$\max\left(\left|\frac{a_{12}}{a_{11}}\right|, \left|\frac{a_{21}}{a_{22}}\right|\right) = \max\left(\frac{1}{12}, \frac{3}{2}\right) = \frac{3}{2}$$

and because of this the sufficient condition for the convergence of the approximation process is not satisfied. But if we put $x = \frac{1}{3}z$ we obtain the system

$$\left.\begin{aligned} 4z + y &= 14 \\ z - 2y &= -1 \end{aligned}\right\} \tag{104}$$

for which

$$\max\left(\left|\frac{a_{12}}{a_{11}}\right|, \left|\frac{a_{21}}{a_{22}}\right|\right) = \max\left(\frac{1}{4}, \frac{1}{2}\right) = \frac{1}{2}$$

This means that system (104) can be solved by the method of successive approximations.

Of course, no one would think of solving such simple systems as (104) by the method of successive approximations. But for systems with a large number of unknowns this method is sometimes very useful. Sufficient conditions for convergence when solving the system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{105}$$

are established in almost the same way as for a system of equations in two unknowns. In fact, the following statement holds:

The process of successive approximations for the system (105) *converges if one of the following conditions is satisfied*:

$$1)\ \max_i \sum_{j=1}^{n}{}' \left|\frac{a_{ij}}{a_{ii}}\right| < 1 \tag{106}$$

$$2)\ \max_j \sum_{i=1}^{n}{}' \left|\frac{a_{ij}}{a_{ii}}\right| < 1 \tag{107}$$

$$3)\ \max_k \sum_{i,j=1}^{n}{}^{(k)} \left|\frac{a_{ij}}{a_{ii}}\right|^2 < 1 \tag{108}$$

The prime in the sums (106) and (107) means that the term for which $i = j$ should be left out. The symbol (k) in the sum (108) means that the terms for which $i = k$ should be left out as well as the terms for which $i = j$. Note that condition (3) must be fulfilled if

$$\sum_{i,j=1}^{n}{}' \left| \frac{a_{ij}}{a_{ii}} \right|^2 < 1 \tag{109}$$

where the prime means that terms for which $i = j$ are omitted. In most books on computing mathematics this condition is presented in the form (109).

As we have already stated before, we arrive at conditions (106)-(108) if we require the mapping

$$x_1 \to \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} x_2 - \ \ldots \ - \frac{a_{1n}}{a_{11}} x_n$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_n \to \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}} x_1 - \frac{a_{n2}}{a_{nn}} x_2 - \ \ldots \ - \frac{a_{n,n-1}}{a_{nn}} x_{n-1}$$

to be a contraction with respect to some distance. Actually, the condition (106) corresponds to the distance between the points $A(x_1, \ldots, x_n)$ and $B(y_1, \ldots, y_n)$

$$r(A, B) = \max\left(|x_1 - y_1|, \ \ldots, \ |x_n - y_n|\right)$$

the condition (107) corresponds to the distance $r(A, B) = \sum_{i=1}^{n} |x_i - y_i|$ and the condition (108) to the distance $r(A, B) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.

As in the case of two variables it is sometimes useful to substitute for the unknowns $x_1, \ldots, x_n$ new unknowns proportional to them: $y_1 = p_1 x_1, \ldots, y_n = p_n x_n$, where $p_1 > 0, \ \ldots, \ p_n > 0$.

In this case the conditions (106), (107), (108) take the form

$$(1')\ \max_i \sum_{j=1}^{n}{}' \left| \frac{a_{ij}}{a_{ii}} \right| \frac{p_i}{p_j} < 1 \tag{106'}$$

$$(2')\ \max_j \sum_{i=1}^{n}{}' \left|\frac{a_{ij}}{a_{ii}}\right| \frac{p_i}{p_j} < 1 \qquad (107')$$

$$(3')\ \max_k \sum_{i,j=1}^{n}{}^{(k)} \left|\frac{a_{ij}}{a_{ii}}\right|^2 \frac{p_i^2}{p_j^2} < 1 \qquad (108')$$

In particular, when $p_i = |a_{ii}|$, these conditions assume the form

$$(1'')\ \max_i \sum_{i=1}^{n}{}' \left|\frac{a_{ij}}{a_{jj}}\right| < 1 \qquad (106'')$$

$$(2'')\ \max_j \sum_{i=1}^{n}{}' \left|\frac{a_{ij}}{a_{jj}}\right| < 1 \qquad (107'')$$

$$(3'')\ \max_k \sum_{i,j=1}^{n}{}^{(k)} \left|\frac{a_{ij}}{a_{jj}}\right|^2 < 1 \qquad (108'')$$

As an example, let us consider the system of equations

$$\begin{aligned} x - 0.06y - 0.5z &= -2.6 \\ -0.2x + y - 0.4z &= 3 \\ -0.1x + 0.5y + z &= 3.9 \end{aligned}$$

The conditions (1), (2), and also (1'), (2') are not satisfied for this system. Similarly, inequality (109) does not hold in this case — the sum of squares of non-diagonal elements is equal to 1.07. But

$$\max_k \sum^{(k)} \left|\frac{a_{ij}}{a_{ii}}\right|^2 = \max(0.6^2 + 0.5^2 + 0.2^2 + 0.4^2;$$

$$0.6^2 + 0.5^2 + 0.1^2 + 0.5^2;\ 0.2^2 + 0.4^2 + 0.1^2 + 0.5^2) = 0.87 < 1$$

and therefore the system can be solved by the method of successive approximations.

Note that all the conditions formulated above are sufficient for the process of successive approximations to converge, but they are by no means necessary. Choosing different definitions of the distance between the points and writing down the contraction condition we arrive at new convergence conditions. However, we do not intend to go into this problem here.

The remarks made in Sec. 5 apply also to systems of linear equations. For instance, the result of the approximations does not depend on the initial approximation. So an error made in the course of the computation does not invalidate the subsequent computations, but only retards the progress towards the final result.

Various forms of the method of successive approximations are used to solve systems of linear equations. Thus, in some methods after the approximate value $x_1^{(n+1)}$ is found, it is substituted together with $x_3^{(n)}$, $x_4^{(n)}, \ldots, x_m^{(n)}$ to find $x_2^{(n+1)}$; then $x_1^{(n+1)}, x_2^{(n+1)}, x_4^{(n)}, \ldots, x_m^{(n)}$ are substituted to find $x_3^{(n+1)}$, and so on. The description of all the possible methods of approximations used for solving systems of linear equations could well form the subject of another book.

27. Successive Approximations in Geometry

We have described the application of the method of successive approximations to the solution of equations and systems of equations. The method is also applied to some problems of geometry, for instance, to the problem of computing the length of the circumference. As is well known, to compute the length of the circumference the practice is to first find the perimeter of the inscribed square, next the perimeter of the inscribed regular octagon, of a regular polygon with 16 sides, and so on. The limit of these perimeters is equal to the length of the circumference. In the process each subsequent perimeter is computed with the aid of the preceding one. This is done in the following way.

Denote the side of a regular 2^n-gon by A_n and its perimeter by P_n. For instance, A_2 is the side of a square and therefore $A_2 = R\sqrt{2}$, $P_2 = 4R\sqrt{2}$. Suppose we have already found P_n. Then obviously

$$A_n = \frac{P_n}{2^n}$$

In geometry it is proved that the side a_{2n} of a regular inscribed 2^n-gon is expressed through the side a_n of a regular inscribed n-gon and the

radius R of the circumference by the formula*

$$a_{2n} = R\sqrt{2 - \sqrt{4 - \frac{a_n^2}{R^2}}} \tag{110}$$

Consequently, the side A_{n+1} of a regular inscribed 2^{n+1}-gon is expressed in terms of the side A_n of a regular 2^n-gon by the formula

$$A_{n+1} = R\sqrt{2 - \sqrt{4 - \frac{A_n^2}{R^2}}}$$

Since $A_n = \frac{P_n}{2^n}$ and $A_{n+1} = \frac{P_{n+1}}{2^{n+1}}$, it follows that

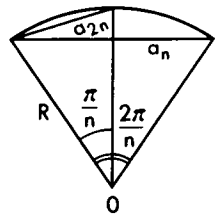
$$P_{n+1} = 2^{n+1} R\sqrt{2 - \sqrt{4 - \frac{P_n^2}{2^{2n}R^2}}} \tag{110'}$$

The sequence of numbers $P_2, P_3, \ldots, P_n, \ldots$ tends to the length of the

*This formula is most easily derived using trigonometry. Obviously, if a_n is the side of a regular inscribed n-gon and a_{2n} the side of a regular inscribed $2n$-gon, then

$$a_n = 2R\sin\frac{\pi}{n} \quad \text{and} \quad a_{2n} = 2R\sin\frac{\pi}{2n}$$

(see drawing). Since



$$\sin\frac{\alpha}{2} = \sqrt{\frac{1 - \cos\alpha}{2}}$$

it follows that

$$a_{2n} = 2R\sin\frac{\pi}{2n} = 2R\sqrt{\frac{1 - \cos\frac{\pi}{n}}{2}} =$$

$$= R\sqrt{2 - 2\sqrt{1 - \sin^2\frac{\pi}{n}}} = R\sqrt{2 - \sqrt{4 - \frac{a_n^2}{R^2}}}$$

circumference, i. e. to the value $2\pi R$. Therefore, formula (110) may be regarded as the formula for computing $2\pi R$ with the aid of the method of successive approximations. Using this method one may find the value of π to any number of decimal places.

There is another method for the approximate computation of π called the *method of equal perimeters*. In this method a regular 2^n-gon is replaced by a regular 2^{n+1}-gon having the same perimeter. Denote the apothem of the regular 2^n-gon by l_n, and the radius of the circumscribed circle by r_n. We denote the apothem of the regular 2^{n+1}-gon having the same perimeter as the 2^n-gon by l_{n+1}, and the radius of the circle circumscribed around it by r_{n+1}.

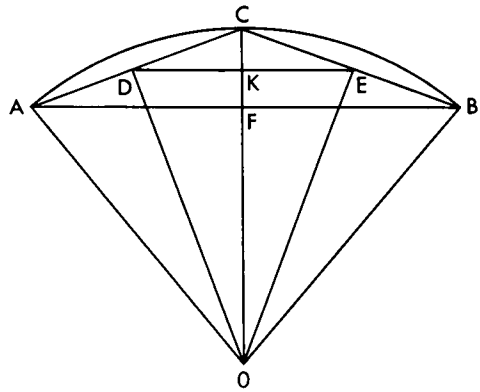


Fig 23

Let AB (Fig. 23) be the side of the 2^n-gon inscribed in the circle of radius r_n. Connect the middle C of the arc AB with points A and B and draw the line DE joining the midpoints D and E of the sides AC and BC, respectively, of the triangle ACB. The angle DOE will, obviously, be equal to half the angle AOB. Therefore DE is a side of the regular 2^{n+1}-gon inscribed in a circle of radius OD. Since $DE = \frac{1}{2}AB$, the perimeter of the 2^{n+1}-gon is equal to the perimeter of the 2^n-gon. This means that $r_{n+1} = OD$, $l_{n+1} = OK$.

It may easily be computed that

$$l_{n+1} = OK = \frac{r_n + l_n}{2} \tag{111}$$

Then we find from the right-angled triangle ODC that

$$r_{n+1} = \sqrt{r_n l_{n+1}} \tag{112}$$

Formulas (111) and (112) express r_{n+1} and l_{n+1} in terms of r_n and l_n.

The perimeters of the polygons do not change with increasing n, and the numbers r_n and l_n approach the same limit. This limit is equal to the radius of the circle whose length is equal to the perimeter of the polygons. If we choose the initial polygon so that its perimeter is equal to 2, then r_n and l_n both approach the number $\frac{1}{\pi}$:

$$\lim_{n\to\infty} r_n = \frac{1}{\pi}, \quad \lim_{n\to\infty} l_n = \frac{1}{\pi}$$

For instance, if we choose for the first polygon a square with the side 1/2, we have $r_2 = \frac{\sqrt{2}}{4}$, $l_2 = \frac{1}{4}$. Therefore the following statement is true: if we put $r_2 = \frac{\sqrt{2}}{4}$, $l_2 = \frac{1}{4}$ and compute the values r_{n+1}, l_{n+1}, $n = 2, 3, \ldots$, using formulas (111) and (112), we would obtain

$$\lim_{n\to\infty} r_n = \lim_{n\to\infty} l_n = \frac{1}{\pi}$$

These formulas may be used to find approximate values of $\frac{1}{\pi}$. In order to do this one should continue the computation until the values of r_n and l_n coincide within the accuracy desired. This common value of r_n and l_n will be the value of $\frac{1}{\pi}$ within the accuracy specified.

28. Conclusion

This book has acquainted us with the applications of the method of successive approximations to various problems: to planning, to the extraction of roots, to the solution of equations, to the computation of the length of a circumference. This by no means exhausts the list of the various applications of this method. A great number of problems lead to differential equations (which contain derivatives of the unknown functions), to integral equations and to equations of a still more com-

plex kind. One of the most powerful methods for the approximate solution of these equations is the method of successive approximations. Of course, its application in such cases is much more complicated than in the case of algebraic equations. But it can be said that if not for the method of successive approximations, not one of the enormous physical and technical problems which are tackled nowadays could be solved. For instance, the method is used in computing the motion of an artificial satellite, in designing an atomic reactor and in research into the structure of the atom. However, a discussion of the applications of the method of successive approximations outside the field of elementary mathematics would be beyond the scope of this book.

Exercises

For the reader to be able to test his memory of the methods of approximate solution of equations discussed in this book we present several examples of approximate solutions of equations.

Solve the following equations using the method of iteration*:

1. $x = \dfrac{1}{(x+1)^2}$

2. $x = (x+1)^3$

3. $x = 4 + \sqrt[3]{\dfrac{x-1}{x+1}}$

4. $x = 2 \pm \sqrt[4]{x}$

5. $x = \sqrt[4]{5-x}$

6. $4 - x = \tan x$

7. $x^2 = \sin x$

8. $x^3 = \sin x$

9. $x = \arcsin \dfrac{x+1}{4}$

10. $x = \cos x$

11. $x = \dfrac{1}{\cos x}$

12. $x = 1 + \dfrac{1}{10} \sin x$

13. $x = \pm \sqrt{\log(x+2)}$

14. $x^2 = \ln(x+1)$

15. $\ln x = 4 - x^2$

16. $\ln x = 2 - x$

---

*In some of the examples the reader must first reduce the equation to the form $x = \varphi(x)$.

17. $x^2 = e^x + 2$

18. $\log x = 0.1x$

19. $\tan x = \log x$

20. $x = \frac{1}{10} e^{-x}$

Solve the following equations using Newton's method:

21. $x^3 - 5x + 1 = 0$

22. $x^3 - 9x^2 + 20x - 11 = 0$

23. $x^3 - 3x^2 - 3x + 11 = 0$

24. $x^5 + 5x + 1 = 0$

25. $\sin x + x = 1$

26. $x^2 - 10 \log x - 3 = 0$

27. Solve with an accuracy of 0 001 the following systems of equations using the method of successive approximations:

a)
$$\begin{cases} x = 0\,2y - 0.1z + 0\,898 \\ y = 0\,3x + 0.15z + 1.383 \\ z = 0.25x - 0.4y + 3.677 \end{cases}$$

b)
$$\begin{cases} x = \frac{1}{4} \sin(x + y) + 0\,336 \\ y = -\frac{1}{4} \sin(x - y) + 0\,362 \end{cases}$$

c)
$$\begin{cases} x = \sqrt{x + 2y} - 0.710 \\ y = \sqrt{y - x} + 1 \end{cases}$$

1. Put $\varphi(x)=\dfrac{1}{(x+1)^2}$. Then $\varphi'(x)=\dfrac{-2}{(1+x)^3}$. We have $\varphi(0)=1>0$, $\varphi(1)=\dfrac{1}{4}<1$. Therefore the interval $[0, 1]$ contains a root of the equation. However, we cannot apply the method of successive approximations to this interval because $|\varphi'(0)|=2>1$. To narrow down the interval note that $\varphi(0.4)=\dfrac{1}{1.96}>0.4$, and so the root of the equation lies in the interval $[0.4, 1]$. If $0.4\leqslant x\leqslant 1$, $|\varphi'(x)|\leqslant\dfrac{2}{1.4^3}<1$ and therefore the method of successive approximations can be applied. Putting $x_1=0.4$ we obtain after 11 approximation steps that $x_{11}\approx\varphi(x_{11})\approx 0.4655$

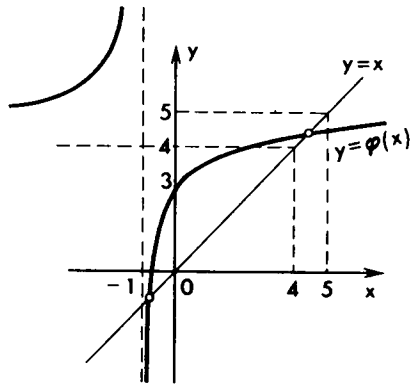


Fig. 24

Therefore with an accuracy of 0.0001 we have $x=0.4655$.

2. Put $\varphi(x)=(x+1)^3$. Then $\varphi'(x)=3(x+1)^2$ and $\varphi(-2)=-1>-2$, $\varphi(-3)=-8<-3$. Therefore the interval $[-3, -2]$ contains the root of this equation. However, we cannot apply the method of successive approximations for in the interval $[-3, -2]$, $|\varphi'(x)|>1$. Rewrite the equation in the form

$$x=\sqrt[3]{x}-1$$

Then $\psi(x)=\sqrt[3]{x}-1$ and $\psi'(x)=\dfrac{1}{3\sqrt[3]{x^2}}$ In the interval $[-3, -2]$ we have $|\psi'(x)\leqslant\dfrac{1}{3\sqrt[3]{x}}<1$, and therefore we can use the method of

successive approximations. Putting $x_1 = -2$, we obtain $x_6 \approx \psi(x_6) \approx$ ≈ -2.325. Therefore with an accuracy of 0.001 we have $x = -2.325$.

3. Put $\varphi(x) = 4 + \sqrt[3]{\dfrac{x-1}{x+1}}$ We have

$$\varphi'(x) = \frac{2}{3\sqrt[3]{(x-1)^2(x+1)^4}}$$

Figure 24 shows that the straight line $y = x$ intersects the curve $y = 4 + \sqrt[3]{\dfrac{x-1}{x+1}}$ in two points lying in the intervals $[-1, 0]$ and $[4, 5]$, respectively. On the interval $[4, 5]$ we have $|\varphi'(x)| \leqslant \dfrac{2}{15\sqrt[3]{45}} < 1$. Putting $x_1 = 4$ we have $x_3 \approx \varphi(x_3) \approx 4.870$. Therefore with an accuracy of 0.001 we have $x_3 = 4.870$.

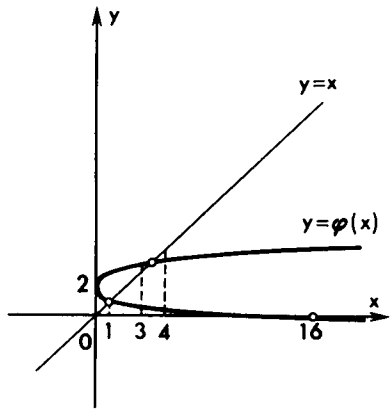


Fig 25

In the interval $[-1, 0]$ the method of successive approximations cannot be employed directly. Rewrite the equation in this section in the form $(x - 4)^3 = \dfrac{x-1}{x+1}$, whence $\dfrac{x-1}{(x-4)^3} = x + 1$ and $x = \dfrac{x-1}{(x-4)^3} - 1$. Here $\psi(x) = \dfrac{x-1}{(x-4)^3} - 1$ and $\psi'(x) = \dfrac{-2x-1}{(x-4)^4}$. Obviously, for $-1 \leqslant x \leqslant 0$ we have $|\psi'(x)| \leqslant \dfrac{1}{256} < 1$ and we may now use the method of successive approximations. Putting $x_1 = 0$, we have $x_2 \approx \psi(x_2) \approx -0.9840$. This means that with an accuracy of 0.0001 we have $x = -0.9840$.

We have found two roots: $x = -0.9840$, $x = 4.870$.

4. Here $\varphi_1(x) = 2 + \sqrt[4]{x}$ and $\varphi_2(x) = 2 - \sqrt[4]{x}$ It may be seen from Fig 25

that the equation $x = 2 + \sqrt[4]{x}$ has a root on the interval [3, 4]. In this interval

$$|\varphi'(x)| = \frac{1}{4\sqrt[4]{x^3}} < 1$$

Apply the method of successive approximations Putting $x_1 = 4$, we have $x_4 \approx \varphi(x_4) \approx 3.353$ Hence, with an accuracy of 0.001 we have $x = 3.353$

Now solve the equation $x = 2 - \sqrt[4]{x}$. Its root is $x = 1$. Hence, the roots of the equation are 1 and 3.353

5. Here $\varphi(x) = \sqrt[3]{5 - x}$ and $\varphi'(x) = \frac{-1}{3\sqrt[3]{(5-x)^2}}$ We have $\varphi(1) = \sqrt[3]{4} < 2$, $\varphi(2) = \sqrt[3]{3} > 1$

Therefore there is a root of the equation on the interval [1, 2]. In this interval $|\varphi'(x)| \leqslant \frac{1}{3\sqrt[3]{9}} < 1$. Putting $x_1 = 1$, we have $x_5 \approx \varphi(x_5) \approx 1.516$ Hence, with an accuracy of 0.001 we have $x = 1.516$.

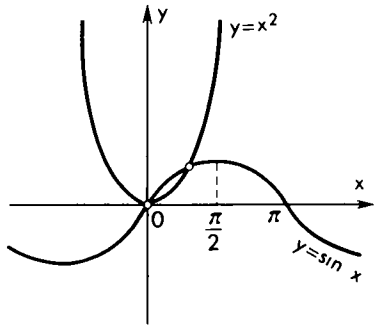


Fig. 26

6. Write the equation in the form

$$x = \arctan(4 - x)$$

Here $\varphi(x) = \arctan(4 - x)$. We have $\varphi(1) = \arctan 3 \approx 1.25$, $\varphi(2) = \arctan 2 \approx$ ≈ 1.10. It follows from this that the equation has a root lying on the interval [1, 2]. In this interval we have

$$|\varphi'(x)| = \frac{1}{1 + (4 - x)^2} \leqslant \frac{1}{5}$$

Therefore the method of successive approximations is applicable Putting $x_1 = 1$, we have $x_4 \approx \varphi(x_4) \approx 1.225$. Therefore with an accuracy of 0.001 we have $x = 1.225$

7. The given equation has a root $x = 0$ It may be seen from Fig. 26 that the second root is positive. Therefore it satisfies the equation $x = \sqrt{\sin x}$. Here

$\varphi(x) = \sqrt{\sin x}$ and $\varphi'(x) = \dfrac{\cos x}{2\sqrt{\sin x}}$ Since

$$\varphi\left(\frac{1}{2}\right) = \sqrt{\sin\frac{1}{2}} \approx \sqrt{0.4794} > \frac{1}{2}$$

and

$$\varphi(1) = \sqrt{\sin 1} \approx \sqrt{0.8414} < 1$$

it follows that the equation has a root in the interval $\left[\frac{1}{2}, 1\right]$ In this interval we have

$$|\varphi'(x)| \leqslant \frac{\cos\frac{1}{2}}{2\sqrt{\sin\frac{1}{2}}} \approx \frac{0.8703}{1.3846} < 1$$

and hence the method of successive approximations converges Putting $x_1 = 1$, we obtain $x_7 \approx \varphi(x_7) \approx 0.8768$ Consequently, with an accuracy of 0.0001, the second root of the equation is 0.8768

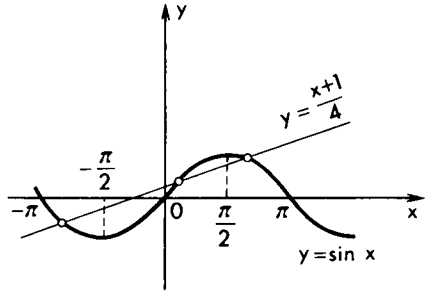


Fig. 27

8. This equation is solved in the same way as the preceding one Rewriting the equation in the form

$$x = \sqrt[3]{\sin x}$$

and putting $x_1 = 1$ we obtain $x_6 \approx \varphi(x_6) \approx 0.9286$. Therefore, with an accuracy of 0.001 one of the roots of the equation is equal to 0.9286 Since both sides of the equation are odd functions, there is also another root equal to -0.9286 The third root is 0.

9. The equation may be rewritten in the form $\dfrac{x+1}{4} = \sin x$, $-\pi/2 \leqslant x \leqslant \pi/2$. It may be seen from Fig. 27 that this equation has a single root lying between 0 and $\pi/2$.

In this case

$$\varphi(x) = \arcsin\frac{x+1}{4}; \quad \varphi'(x) = \frac{1}{\sqrt{16-(x+1)^2}}$$

On the interval $[0, \pi/2]$ we have $|\varphi'(x)| < \frac{1}{\sqrt{3}} < 1$. Putting $x_1 = 0$, we find $x_8 \approx \varphi(x_8) \approx 0.3422$, and therefore with an accuracy of 0.0001 we have $x = 0.3422$.

10. Since $\cos 0 = 1$, $\cos 1 > 0$, the equation $x = \cos x$ has a root on the interval $[0, 1]$. Since $|\varphi'(x)| \leqslant \sin 1 < 1$, the method of successive approximations may be used. Putting $x_1 = 1$, we get $x = 0.7391$ with an accuracy of 0.0001

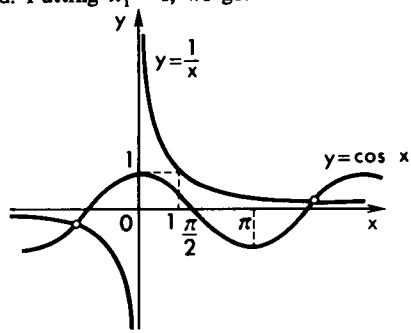


Fig 28

11. It may be seen from Fig. 28 that the positive roots of the equation lie close to the points of intersection of the graph of the function $y = \cos x$ with the x-axis and are to the right of the points of intersection of the type $\frac{\pi}{2} + (2k+1)\pi$ and to the left of the points of intersection of the type $\frac{\pi}{2} + 2k\pi$.

To find the solution in the vicinity of the point $x = \frac{\pi}{2} + n\pi$ put $x - n\pi - \frac{\pi}{2} = y$. The equation will assume the form

$$y + n\pi + \frac{\pi}{2} = \frac{1}{\cos\left(y + n\pi + \frac{\pi}{2}\right)} = \frac{(-1)^{n+1}}{\sin y}$$

Since $-\frac{\pi}{2} \leqslant y \leqslant \frac{\pi}{2}$, the equation may be written thus:

$$y = (-1)^{n+1}\arcsin\frac{1}{y + n\pi + \frac{\pi}{2}}$$

Here

$$\varphi(y) = (-1)^{n+1} \arcsin \frac{1}{y + n\pi + \frac{\pi}{2}}$$

and

$$\varphi'(y) = \frac{(-1)^n}{\sqrt{\left(y + n\pi + \frac{\pi}{2}\right)^2 - 1}\left(y + n\pi + \frac{\pi}{2}\right)}$$

Clearly, in the vicinity of point $y = 0$, we have $|\varphi'(y)| < q < 1$, and so we may use the method of successive approximations. Find the solution for $n = 1$ with an accuracy of 0.001. Let $y_0 = 0$. Then $y_2 \approx \varphi(y_2) \approx 0.204$. Therefore $y \approx 0.204$. Hence $x = \frac{3}{2}\pi + y \approx 4.917$.

To find the first negative root, put $n = -1$. We shall obtain the equation

$$y = \arcsin \frac{1}{y - \frac{\pi}{2}}$$

Put $y_0 = 0$. Then

$$y_6 \approx \varphi(y_6) \approx -0.503$$

Hence $y \approx -0.503$ and so $x \approx -2.074$.

For large values of $|n|$ the method of successive approximations gives an approximate formula for y:

$$y \approx \varphi(y_0) = (-1)^{n+1} \arcsin \frac{1}{n\pi + \frac{\pi}{2}} \approx \frac{(-1)^{n+1} \times 2}{(2n+1)\pi}$$

Therefore

$$x \approx \frac{\pi}{2}(2n+1) + \frac{(-1)^{n+1} \times 2}{(2n+1)\pi}$$

12. Putting $x_1 = 0$ we obtain $x_3 \approx \varphi(x_3) \approx 1.088$. Therefore with an accuracy of 0.001 we have $x = 1.088$.

13. First solve the equation $x = \sqrt{\log(x+2)}$. We have $\varphi(x) = \sqrt{\log(x+2)}$ and therefore

$$\varphi'(x) = \frac{\log e}{2(x+2)\sqrt{\log(x+2)}}$$

Since $\varphi(0) = \sqrt{\log 2} > 0$, $\varphi(1) = \sqrt{\log 3} < 1$, the equation has a root on the interval $[0, 1]$. On this interval the inequality $|\varphi'(x)| < q < 1$ holds. Therefore the method of successive approximations is applicable. Putting $x_1 = 1$ we have

$x_5 \approx \varphi(x_5) \approx 0.6507$. Therefore the root of the equation $x = \sqrt{\log(x+2)}$ is equal to $x = 0\,6507$ with an accuracy of 0.0001

Consider the equation

$$x = -\sqrt{\log(x+2)}$$

Here

$$\varphi(x) = -\sqrt{\log(x+2)}$$

Since $\varphi(0) = -\sqrt{\log 2} = -0.55$, $\varphi\left(-\frac{1}{2}\right) = -\sqrt{\log 1.5} = -0.42$, the equation has a root on the interval $\left[-\frac{1}{2}, 0\right]$. Putting $x_1 = 0$ we find $x_8 \approx$ $\approx \varphi(x_8) \approx -0\,4397$. Hence, with an accuracy of 0 0001 we have $x = -0\,4397$

14. One of the roots of the equation is $x = 0$ To find another root, write the equation in the form $x = \pm\sqrt{\ln(x+1)}$. For the equation $x = \sqrt{\ln(x+1)}$ we have

$$\varphi\left(\frac{1}{2}\right) = \sqrt{\ln\frac{3}{2}} > \frac{1}{2}$$

$$\varphi(1) = \sqrt{\ln 2} < 1$$

Hence the equation has a root on the interval [1/2, 1] Since $\varphi'(x) = -\frac{1}{2(x+1)\sqrt{\ln(x+1)}}$ we have $|\varphi'(x)| < q < 1$ on the interval [1/2, 1]. Put $x_1 = 1$, then $x_9 \approx \varphi(x_9) \approx 0\,7469$ Hence, with an accuracy of 0 0001 we have $x = 0.7469$ The equation $x = -\sqrt{\ln(x+1)}$ has no roots other than $x = 0$. Hence $x = 0$ or $x = 0.7469$.

15. Rewrite the equation in the form

$$x = \sqrt{4 - \ln x}$$

Here

$$\varphi(1) = 2, \quad \varphi(2) = \sqrt{4 - \ln 2}, \quad \varphi'(x) = \frac{-1}{2x\sqrt{4 - \ln x}}$$

Since $\varphi(1) = 2$, $\varphi(2) = \sqrt{4 - \ln 2} < 2$, the equation has a root on the interval [1, 2]. It is evident from Fig. 29 that there are no other roots. Putting $x_1 = 2$ we obtain

$$x_4 \approx \varphi(x_4) \approx 1\,841$$

Hence $x = 1.841$ with an accuracy of 0.001.

16. Write the equation in the form

$$x = 2 - \ln x$$

Here $\varphi(x) = 2 - \ln x$, $\varphi'(x) = -\frac{1}{x}$ It may be seen from Fig 30 that the root

of the equation lies on the interval $[1, 2]$ In this interval $|\varphi'(x)| \leqslant 1$ Putting $x_1 = 1\,5$ we obtain $x_{13} \approx \varphi(x_{13}) \approx 1\,557$. Hence, $x = 1\,557$ with an accuracy of 0 001

17. It may be seen from Fig 31 that the equation has only one negative root Write the equation in the form

$$x = -\sqrt{e^x + 2}$$

Then

$$\varphi(x) = -\sqrt{e^x + 2}, \quad \varphi'(x) = \frac{-e^x}{2\sqrt{e^x + 2}}$$

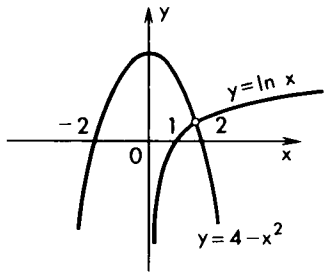


Fig 29

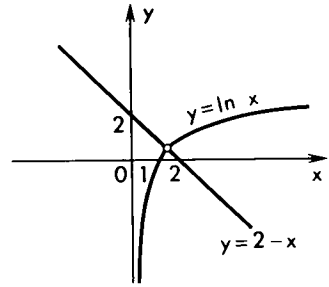


Fig 30

and

$$\varphi(-1) = -\sqrt{2 + e^{-1}} \approx -1\,54; \quad \varphi(-2) = -\sqrt{2 + e^{-2}} \approx -1.46$$

Therefore the root lies on the interval $[-2, -1]$ Putting $x_1 = -1$ we get $x_4 \approx \varphi(x_4) \approx -1.492$. Hence with an accuracy of 0.001 we have $x = -1\,492$

18. Clearly one of the roots of the equation is $x_1 = 10$. To find the second root, write the equation in the form $x = 10^{0\,1x}$. Here $\varphi(x) = 10^{0\,1x}$, $\varphi'(x) = 0.1 \times 10^{0\,1x} \ln 10$ Also $\varphi(1) = 10^{0\,1} > 1$, $\varphi(2) = 10^{0\,2} < 2$ Therefore the equation has a root on the interval $[1, 2]$. In this interval $|\varphi'(x)| \leqslant 0.1 \times 10^{0,2} \ln 10 \approx 0\,37 < 1$. Now the method of successive approximations can be used. Putting $x_1 = 2$ we find $x_7 \approx \varphi(x_7) \approx 1.372$. Hence, with an accuracy of 0.001 we have $x = 1.372$

19. It may be seen from Fig. 32 that the equation has a root in each of the intervals $\left[\frac{\pi}{2} + n\pi, \frac{\pi}{2} + (n+1)\pi\right]$, $n = 0, 1, \ldots$, with the roots lying in the right halves of the sections. To find the first positive root, substitute $x = \frac{3\pi}{2} - y$. The equation will assume the form

$$\cot y = \log\left(\frac{3\pi}{2} - y\right)$$

from which we get

$$y = \operatorname{arccot}\left[\log\left(\frac{3}{2}\pi - y\right)\right]$$

since $0 < y < \pi$. Here

$$\varphi(y) = \operatorname{arccot}\left[\log\left(\frac{3}{2}\pi - y\right)\right]$$

and

$$\varphi'(y) = \frac{-\log e}{\left[1 + \log^2\left(\frac{3}{2}\pi - y\right)\right]\left(\frac{3}{2}\pi - y\right)}$$

On the interval $[0, \pi]$ there is a root of our equation. Moreover, in this interval $|\varphi'(y)| < 1$. Apply the method of successive approximations. Putting $y_1 = 0$ we have $y_4 \approx \varphi(y_4) \approx 1.059$. Hence with an accuracy of 0.001 we have $y = 1.059$ and therefore $x = 3.654$.

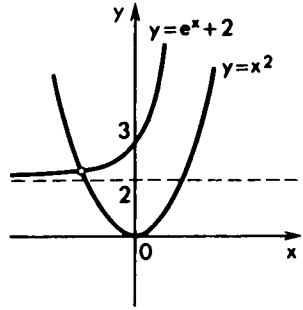


Fig. 31

To find the second positive root we put $x = \frac{5}{2}\pi - y$. The equation assumes the form

$$y = \operatorname{arccot}\left[\log\left(\frac{5}{2}\pi - y\right)\right]$$

Putting $y_1 = 0$ we get $y_4 \approx \varphi(y_4) \approx 0.870$. Hence with an accuracy of 0.001 we have $y = 0.870$ and $x = 6.984$.

20. It may be seen from Fig. 33 that the equation has a single root lying between 0 and 1. We have $\varphi(x) = \frac{1}{10}e^{-x}$, $\varphi'(x) = -\frac{1}{10}e^{-x}$. On the interval $[0, 1]$ the inequality $|\varphi'(x)| \leqslant \frac{1}{10}$ holds, which makes the use of the method of

successive approximations possible. Putting $x_1 = 0$ we have $x_4 \approx \varphi(x_4) \approx$ ≈ 0.091. Hence with an accuracy of 0,0001 we have $x = 0.091$.

21. Let

$$f(x) = x^3 - 5x + 1$$

Then

$$f'(x) = 3x^2 - 5, \quad f''(x) = 6x$$

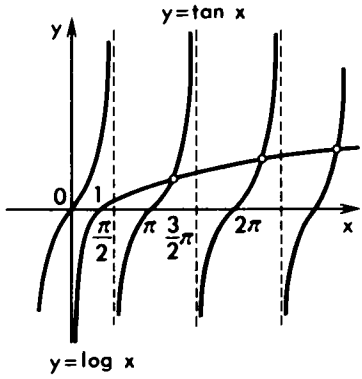


Fig 32

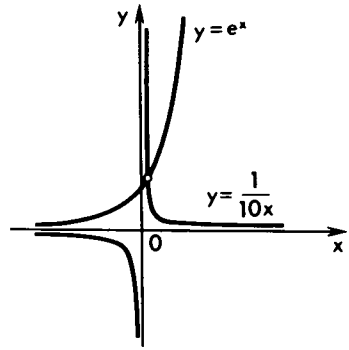


Fig 33

Using Newton's formula we have

$$\beta_{n+1} = \beta_n - \frac{\beta_n^3 - 5\beta_n + 1}{3\beta_n^2 - 5}$$

Compute a table of the values of the function:

x	-3	-2	-1	0	1	2	3
$f(x)$	-11	3	5	1	-3	-1	13

It is seen from this table that the equation $x^3 - 5x + 1 = 0$ has roots on the intervals $[-3, -2]$, $[0, 1]$, $[2, 3]$.

First let us find the root lying on the interval $[-3, -2]$. Since in this interval $f''(x) < 0$, we choose the initial value $\beta_0 = -3$ (because $f(\beta_0) = -11$ is a negative number). We have

$$\beta_1 = -3 - \frac{(-3)^3 - 5(-3) + 1}{3(-3)^2 - 5} = -2.5$$

Continuing the computation we find that $\beta_3 \approx \beta_4 \approx -2.331$ and therefore with an accuracy of 0.001 the root of the equation on the interval $[-3, -2]$ is -2.331.

Next we find the root lying on the interval $[0, 1]$. Here we have $f''(x) \geqslant 0$. Therefore we put $\beta_0 = 0$ Hence we obtain

$$\beta_1 = 0 - \frac{0^3 - 5 \times 0 + 1}{3 \times 0^2 - 5} = 0{,}2, \quad \beta_3 \approx \beta_4 \approx 0.202.$$

With an accuracy of 0 001 we have $x = 0.202$.

Finally, to find the root on the interval $[2, 3]$ we put $\beta_0 = 3$ and get

$$\beta_1 = 3 - \frac{3^3 - 5 \times 3 + 1}{3 \times 3^2 - 5} \approx 2.409$$

Continuing the computation we find $\beta_4 \approx \beta_5 \approx 2.128$ Hence with an accuracy of 0.001 the root is equal to 2.128. We have found three roots: $x_1 = -2.331$; $x_2 = 0.202$; $x_3 = 2.128$.

Solve this equation using the improved method of chords. On the interval $[-3, -2]$ we find

$$\alpha_1 = -3 - f(-3)\frac{-3-(-2)}{f(-3)-f(-2)} = -3 + 11\frac{-3}{-14} \approx -2214$$

Since on this interval $f''(x) < 0$, the curve is concave down and we find α_2 using the formula

$$\alpha_2 = -3 - f(-3)\frac{-3-(-2.214)}{f(-3)-f(-2.214)} \approx -2.293$$

Next we find

$$\alpha_3 = -2\,293 - f(-2\,293)\frac{-2\,293 + 2\,214}{f(-2.293) - f(-2\,214)} \approx -2\,331$$

This coincides within the accuracy of 0 001 with the value of x obtained above

The same method is used to solve the equation on the intervals $[0, 1]$ and $[2, 3]$.

22. Here we have

$$f(x) = x^3 - 9x^2 + 20x - 11$$

$$f'(x) = 3x^2 - 18x + 20$$

$$f''(x) = 6x - 18 = 6(x - 3)$$

Compile the table of the values of the function $f(x)$:

x	0	1	2	3	4	5	6
$f(x)$	-11	1	1	-5	-11	-11	1

The roots of the equation lie on the intervals [0, 1], [2, 3], [5, 6].

On the interval [0, 1] we put $\beta_0 = 0$ and $\beta_4 \approx \beta_5 \approx 0\,834$. On the interval [2, 3] we put $\beta_0 = 3$ and find $\beta_2 \approx \beta_3 \approx 2.216$ On the interval [5, 6] we put $\beta_1 = 0$ and find $\beta_4 \approx \beta_5 \approx 5.249$. We have found (with an accuracy of 0.001) three roots of the equation:

$$x_1 = 0.834, \quad x_2 = 2.216, \quad x_3 = 5.249$$

23. Here $f(x) = x^3 - 3x^2 - 3x + 11, f'(x) = 3x^2 - 6x - 3, f''(x) = 6x - 6 = 6(x - 1)$. Compile a table of the values of $f(x)$

x	-2	-1	0	1	2	3
$f(x)$	-3	10	11	6	1	2

The equation has one real root lying on the interval $[-2, -1]$. To find this root we put $\beta_0 = -2$. We obtain $\beta_2 \approx \beta_3 \approx -1.847$. Therefore with an accuracy of 0.001 we have $x = -1.847$.

24. Here $f(x) = x^5 + 5x + 1, f'(x) = 5x^4 + 5, f''(x) = 20x^3$. The table of the $f(x)$ values is

x	-1	0	1
$f(x)$	-5	1	7

Therefore the equation has a root on the interval $[-1\,0]$ We put $\beta_0 = -1$. With an accuracy of 0.0001 we have $\beta_3 \approx \beta_4 \approx -0\,1999$. Therefore with the accuracy specified $x = -0.1999$.

25. Here $f(x) = \sin x + x - 1, f'(x) = \cos x + 1, f''(x) = -\sin x$. A table of the values of $f(x)$ is

x	0	1	2
$f(x)$	-1	0 8115	1 9093

The root lies on the interval [0, 1] Putting $\beta_0 = 0$ we find $\beta_2 \approx \beta_3 \approx 0\,5110$. Therefore with an accuracy 0.0001 we have $x = 0.5110$.

26. Here $f(x) = x^2 - 10\log x - 3$, $f'(x) = 2x - \dfrac{10}{x \ln 10}$, $f''(x) = 2 + \dfrac{10}{x^2 \ln 10}$. A table of the values of $f(x)$ is

x	0.5	1	2	3
$f(x)$	0.26	-2	-2.01	1.23

The roots of the equation lie on the intervals [0.5, 1] and [2, 3]. On the interval [0.5, 1] we put $\beta_0 = 0.5$ and obtain $\beta_2 \approx \beta_3 \approx 0.535$. Therefore the corresponding root with an accuracy of 0.001 is equal to 0.535. On the interval [2, 3] we put $\beta_0 = 3$ and find $\beta_2 \approx \beta_3 \approx 2.705$.

The equation has two roots: $x_1 = 0.535$; $x_2 = 2.705$.

27. In the system (a) we put $x_0 = 0$, $y_0 = 0$, $z_0 = 0$ and after a few approximations find with an accuracy of 0.001 that $x = 1.021$, $y = 2.150$, $z = 3.072$.

In system (b) we put $x_0 = 0$, $y_0 = 0$ and after a few approximations find (with an accuracy of 0.001) that $x = 0.520$, $y = 0.310$.

In system (c) we put $x_0 = 0$, $y_0 = 0$ and find $x = 1.000$, $y = 2.000$.

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book.
We would also be pleased to receive any other suggestions you may wish to make.
Our address is: USSR, 129820, Moscow, I-110, GSP, Pervy Rizhsky Pereulok, 2 Mir Publishers

Printed in the Union of Soviet Socialist Republics